# MASTER'S THESIS

## De Ontwikkeling van een Vragenlijst voor de Beoordeling van Hoorcolleges in Academisch-Medisch Onderwijs door Collega-docenten.

Turk, Rob

**Award date:**
2020

# De Ontwikkeling van een Vragenlijst voor de Beoordeling van Hoorcolleges in Academisch-Medisch Onderwijs door Collega-docenten

# Constructing a Questionnaire for Peer Reviewing Lectures in Academic Medical Teaching

R.F.M. Turk, MD, MSc

Master Onderwijswetenschappen
Open Universiteit

Datum : 27-02-2020
Begeleiding : Dr. J. Gijselaers

**Inhoud**

# De ontwikkeling van een vragenlijst voor de beoordeling van hoorcolleges in academisch-medisch onderwijs door collega-docenten

R.F.M. Turk, MD, MSc

## Samenvatting

*Achtergrond*. De kwaliteit van een hoorcollege wordt vaak gemeten aan de hand van een studentenevaluatie. Een studentenevaluatie geeft echter geen volledig betrouwbaar beeld van de kwaliteit van het hoorcollege. Het beoordelen van elkaars onderwijsprestaties, peer review of teaching, is een aanvullend instrument om de kwaliteit van een hoorcollege aanvullend te kunnen beoordelen. Objectieve criteria voor het geven van een peer review van academisch-medische hoorcolleges zijn echter schaars. *Doel*. Het ontwikkelen van een wetenschappelijk onderbouwde, korte vragenlijst om het geven van peer review van hoorcolleges binnen de academisch-medische opleiding van ons universitair medisch centrum te ondersteunen. *Procedure en ontwerp*. Een literatuuronderzoek is uitgevoerd binnen ERIC (OVID), PsycINFO (OVID) en Pubmed naar wetenschappelijk onderbouwde criteria en vragen voor de beoordeling van medisch onderwijs. Op basis van de meest voorkomende vragen werd in een iteratief proces met vier professionals (leeftijd 49-66, twee docenten, een onderwijskundige en een docent/onderwijskundige in opleiding, drie vrouwen en een man) een initiële vragenlijst ontwikkeld. Deze vragenlijst werd tijdens hoorcolleges getest in de docentengroep van ons academisch medisch centrum (leeftijd 25-66 jaar, ongeveer 40% man, 60% vrouw). De vragenlijst werd gevalideerd met behulp van SPSS, Principle Axis Factoring, Direct Oblimin. Na analyse werd de vragenlijst door vier andere professionals (leeftijd 40-59, twee docenten, twee onderwijskundigen, twee vrouwen en twee mannen) in consensus aangepast tot een gewenste compacte vragenlijst van 10 vragen, en werden mogelijke constructen opgesteld op basis van de gevonden factoren. *Resultaten*. Uit de 386 gevonden wetenschappelijke artikelen werden 32 artikelen geselecteerd. Een kader van 10 categorieën voor de indeling van de gevonden vragen werd ontwikkeld op basis van acht wetenschappelijke studies. In de literatuur werden 75 unieke vragen gevonden voor de beoordeling van de kwaliteit van onderwijs, waarvan 25 vragen het onderwijsmoment zelf betroffen. Voor de initiële vragenlijst werden 21 vragen geselecteerd. Voor validatie werden 387 vragenlijsten ingevuld door 195 unieke collega-docenten (leeftijd 25-66 jaar, 40% man, 60% vrouw) die het onderwijs beoordeelden van 37 unieke docenten (leeftijd 29-62 jaar, 40% man, 60% vrouw) op 37 unieke onderwijsmomenten binnen ons academisch medisch centrum. In de ontwikkelde vragenlijst konden drie factoren worden berekend: (a) context van het hoorcollege (eigenwaarde 1,0, verklaarde variantie 10,3%, Cronbach's α ,61); (b) leeromgeving (eigenwaarde 1,16, verklaarde variantie 11,6%, Cronbach's α ,67); en (c) studentgerichtheid (eigenwaarde 3,90, verklaarde variantie 39,0%, Cronbach's α ,78). De drie factoren verklaarden 60,9% van de variantie van de vragenlijst. *Conclusie*. In een eerste aanzet tot de ontwikkeling van een vragenlijst voor de ondersteuning van peer review van kwaliteit van hoorcolleges van academisch-medisch onderwijs ontwikkelden wij een korte vragenlijst.

Verder onderzoek is noodzakelijk om de uiteindelijke vragenlijst opnieuw te valideren en onderzoek te verrichten naar inter-rater reliability.
Sleutelwoorden: onderwijsontwikkeling, peer review, vragenlijst, academisch medisch onderwijs, docentprofessionalisering

# Constructing a questionnaire for peer reviewing lectures in academic medical teaching

R.F.M. Turk, MD, MSc

## Summary

*Background*. The quality of a lecture is often examined by a students' evaluation. Students' evaluations, however, are biased and not reliable enough to measure the quality of the lecture completely. Peer review of teaching, judging the quality of teaching by peers, is one of the supplementary methods to evaluate the quality of lectures. Objective criteria for peer reviewing academic medical lectures, however, are scarce. *Aim*. To support the peer review of lectures within the medical education of our Dutch academic medical centre, we aimed to construct a short, evidence-based and validated questionnaire. *Procedure and design*. A literature search was performed in ERIC (OVID), PsycINFO (OVID) and PubMed on evidence based criteria and questions for assessing medical teaching. We selected the questions regarding the moment of instruction only for further use. In an iterative process with four academic professionals (age 49-66, two lecturers, one educational expert and one lecturer/education expert in training, three women, one man) an initial questionnaire for peer reviewing medical teaching was assembled from the questions with the highest prevalence. The questionnaire was tested within the group of lecturers at our academic medical centre (age 25-66, approximately 40% male, 60% female). The questionnaire was validated using SPSS, Principle Axis Factoring, Direct Oblimin. After analysis, four other professionals (age 40-59, two lecturers, two educational experts, two women, two men) customised the questionnaire in consensus to 10 questions, and formulated constructs on the factors found. *Results*. Out of 386 articles, 32 were included. We constructed a framework of 10 categories for classifying the questions, using eight evidence based studies. In literature, 75 unique questions assessing the quality of teaching were found in 12 studies, of which 25 questions reflected the moment of instruction. Twenty-one questions were selected for an initial questionnaire. In validating this questionnaire 387 forms were filled in by 195 unique peers (age 25-66, 40% male, 60% female), reviewing 37 unique lecturers (age 29-62, 40% male, 60% female) in academic medical teaching. Three factors were found within the final questionnaire: (a) context of the lecture (eigenvalue 1.00, explained variance 10.3%, Cronbach's α .61); (b) learning environment (eigenvalue 1.16, explained variance 11.6%, Cronbach's α .67); and (c) student focus (eigenvalue 3.90, explained variance 39.0%, Cronbach's α .78). The three factors explained 60.9% of the variance of the questionnaire. *Conclusion*. In order to construct a questionnaire to support peer reviewing the quality of lectures in academic medical teaching, we constructed a validated short questionnaire. Further research is needed revalidate the final questionnaire and to examine the inter-rater reliability.

Keywords: faculty development, peer review, questionnaire, academic medical teaching, professionalisation of lecturers

# De ontwikkeling van een vragenlijst voor de beoordeling van hoorcolleges in academisch-medisch onderwijs door collega-docenten

R.F.M. Turk, MD, MSc

## Probleemschets en Doel van het Onderzoek

Hoorcolleges binnen een academisch-medische opleiding zijn vaak afwijkend van hoorcolleges binnen andere academische opleidingen doordat bij medische hoorcolleges patiënten aanwezig (kunnen) zijn waarmee interactie plaatsvindt met de studenten (Baldwin, Chandran, & Gusic, 2011). Binnen ons Nederlands Universitair Medisch Centrum (UMC) wordt de kwaliteit van hoorcolleges voornamelijk gemeten door middel van studentenevaluaties, maar deze studentenevaluaties zijn inhoudelijk vooral een evaluatie op themaniveau. Binnen de studentenevaluaties is er weinig ruimte om een specifiek hoorcollege nader te onderzoeken op kwaliteit. Daarnaast worden de uitkomsten van studentenevaluaties beïnvloed door niet-onderwijsgebonden factoren (Kamran, Zibaei, Mirkaimi, & Shahnazi, 2012). De docentengroep van ons UMC, vertegenwoordigd door de groep van *Principal Educators (PE's)*, docenten die door de Raad van Bestuur als uitmuntend en vooruitstrevend zijn benoemd, heeft daarom de behoefte uitgesproken om de kwaliteit van hoorcolleges aanvullend te kunnen meten door middel van regelmatige *peer review of teaching*, het beoordelen van het onderwijs door een gelijke.

Omdat peer review of teaching beïnvloed kan worden door verschillen in visie op onderwijs (Alabi & Weare, 2014; Newman, Lown, Jones, Johansson, & Schwartzstein, 2009) wil de groep van docenten van ons UMC een vragenlijst hebben waarmee het proces van peer review van kwaliteit van hoorcolleges kan worden ondersteund. De binnen ons UMC gebruikte vragenlijst voor de Basis Kwalificatie Onderwijs (BKO), waarin ook peer review wordt toegepast, bestaat uit 23 vragen en wordt door de docenten als te lang ervaren. Onderzoek naar een andere vragenlijst voor dit doel in de bestaande literatuur leverde uitsluitend een vragenlijst van Newman et al. (2009) op. Mede door grote onderlinge verschillen in interpretatie van de verschillende vragen tussen de beoordelaars (Newman et al., 2009), een vijftal tweeledig gestelde vragen, en het feit dat deze vragenlijst ontworpen is voor de Amerikaanse context, wordt de bruikbaarheid van deze vragenlijst binnen ons UMC door de groep van PE's als beperkt gezien. Om deze reden heeft de groep van PE's in 2016 bij de Raad van Bestuur het verzoek ingediend om onderzoek te mogen doen naar de ontwikkeling van een nieuwe vragenlijst voor de ondersteuning van peer review van kwaliteit van hoorcolleges. Deze scriptie beschrijft de eerste fase van de totstandkoming van deze vragenlijst, waarin de vragenlijst zal worden ontwikkeld en worden onderzocht op factoren en interne betrouwbaarheid.

Het doel van het onderzoek van deze scriptie is te komen tot een korte, gevalideerde vragenlijst van maximaal 10 items, voor het ondersteunen van een vrijwillige, formatieve peer review van de kwaliteit van hoorcolleges binnen het academisch-medisch onderwijs van ons UMC. Op verzoek van de groep van docenten staat het toekomstig gebruik van de vragenlijst in ieder geval de eerste jaren

nog los van de binnen ons UMC aanwezige kwaliteitskaders om een onafhankelijke ontwikkeling van de vragenlijst mogelijk te maken en docenten te laten wennen aan het ontvangen en verzorgen van een regelmatige peer review op de kwaliteit van de hoorcolleges. De uiteindelijke vragenlijst zal de eerste jaren daarmee nog geen formele status krijgen binnen de kwaliteitskaders van ons UMC.

**Theoretisch Kader**

Onderwijs wordt door Driscoll (2014, p. 23) gedefinieerd als elke doelbewuste rangschikking van gebeurtenissen die een leerling ondersteunen om de gestelde doelen (op het gebied van bijvoorbeeld kennis, kunde, houding en strategieën) te laten bereiken. Het medisch georiënteerde beroepsonderwijs onderscheidt zich hierin van andere vormen van onderwijs doordat er sprake is van een leeromgeving met patiëntencontacten (Baldwin et al., 2011), en een periode van langdurig en expliciet leren tijdens de uitvoering van werkzaamheden in de opleidingsperiode (Molenaar et al., 2009). Daarnaast onderscheidt de medisch georiënteerde beroepsopleiding zich doordat hoogopgeleide professionals uit de zorg verantwoordelijk zijn voor het onderwijsproces gedurende de opleidingsperiode (Molenaar et al., 2009). Medisch onderwijs mag van hoge kwaliteit worden genoemd wanneer het zich niet alleen richt op het bestaande curriculum, maar ook op de behoeften van studenten, en het studenten stimuleert om actief deel te nemen aan het onderwijs om daarmee hun leerprestaties te vergroten (Baldwin et al., 2011).

Ieder instituut voor hoger onderwijs in Nederland is verplicht de onderwijskwaliteit van de eigen opleiding te evalueren (NVAO, 2018; Wet op het Hoger onderwijs en Wetenschappelijk onderzoek, 1992). Probleem is echter, dat het begrip onderwijskwaliteit geen eenduidige definitie kent en onder meer afhankelijk is van wie de kwaliteit beoordeelt, het kader en de achtergrond van de beoordelaar, eventueel bestaande richtlijnen voor de beoordeling, en de situatie die beoordeeld wordt. Mede daardoor is het begrip onderwijskwaliteit op vele manieren te interpreteren en te meten (Harvey & Green, 1993). De Onderwijsraad (2015) adviseert dan ook om, rekening houdend met de vereisten voor accreditatie, het begrip onderwijskwaliteit breed te interpreteren en lokaal in te vullen en te meten.

Een van de manieren om onderwijskwaliteit te onderzoeken is het evalueren van de effectiviteit van het onderwijs (Harvey & Green, 1993). *Effectiviteit van onderwijs* kan worden gedefinieerd als een begrip dat “should have something to do with getting students underway (marketing questioning), and the advances students make by being underway (lines of questioning, positions arrived at)” (Hill & Herche, 2001, p. 20). Scheerens, Luyten, van Ravens en van Ravens (2010) definiëren effectiviteit van onderwijs als de causale relatie tussen input, processen en context enerzijds en opbrengsten anderzijds. Het begrip effectiviteit kan worden samengesteld en gewogen op basis van vele factoren, waarbij de kwaliteit van de instructie in de regel een van de basiscriteria is voor de beoordeling van de effectiviteit op docentniveau (Scheerens & Blömeke, 2016).

Binnen het hoger onderwijs wordt de kwaliteit van de instructie onder meer onderzocht door het inzetten van studentenevaluaties (Berk, 2013). Onderzoek naar deze studentenevaluaties laat echter zien dat de variantie in de uitkomsten beïnvloed kan worden door niet onderwijs gebonden factoren (Kamran et al., 2012; Rannelli, Coderre, Paget, Woloschuk, Wright, & McLaughlin, 2014). Daarnaast is er twijfel over de deskundigheid onder studenten om bepaalde indicatoren voor de kwaliteit van onderwijs goed te kunnen beoordelen (Berk, 2013). Ook beoordelen studenten het onderwijs vaak anders dan een onderwijskundige dit zou doen (Pettit, Axelson, Ferguson, & Rosenbaum, 2015). In een meta-analyse van onderzoeken wordt dan ook geen correlatie gevonden tussen de uitkomsten van studentenevaluaties en de mate van effectiviteit van onderwijs (Uttl, White, & Gonzalez., 2017).

Voor de beoordeling van de kwaliteit van de instructie is het dan ook noodzakelijk dat ook anderen een mening kunnen geven over het onderwijs (Berk, 2013; Van Note Chism, 2007). Probleem daarbij kan echter wel zijn dat verschillende observatoren ieder een andere mening kunnen hebben over wat de kwaliteit van instructie beïnvloedt (Simendinger et al., 2017; Van Note Chism, 2007). De observatoren moeten dan eerst consensus hebben bereikt over de definitie van de kwaliteit van instructie om tot een goed gewogen oordeel te kunnen komen. Een zekere training of begeleiding in de beoordeling van de kwaliteit van instructie kan hierbij noodzakelijk zijn (Finn & Garner, 2011; Newman et al., 2009). Om te kunnen komen tot deze geadviseerde multidisciplinaire beoordeling zijn aanvullende methoden noodzakelijk waarmee, naast de studentenevaluaties, de kwaliteit van instructie kan worden beoordeeld (Berk, 2013; Van Note Chism, 2007). Het laten beoordelen van het onderwijs door collega-docenten, de zogenaamde *peer review van onderwijs,* is een van de aanvullende methoden om de kwaliteit van de instructie te kunnen beoordelen (Van Note Chism, 2007).

Volgens de definitie van Van Note Chism (2007) is peer review van onderwijs een vooraf besproken, oordelende reflectie door collega's op de onderwijsprestaties van een docent. Peer review kan dienen ter ondersteuning van de verbetering van onderwijs (de formatieve variant), of voor het kunnen maken van personele beslissingen (de summatieve variant). Formatieve peer review is vertrouwelijk en persoonlijk, bedoeld voor het vergroten van de zelfkennis over het verzorgde onderwijs (Van Note Chism, 2007, pp. 3-5). Het geven van een peer review van onderwijs houdt meer in dan alleen het beoordelen van het onderwijsmoment zelf. Ook het beoordelen van de context en de doelen van het onderwijs, de tijdens het onderwijs gebruikte materialen en onderwijsmethoden, en de gebruikte toets- en evaluatiemethoden kunnen onderwerp zijn van een peer review (Van Note Chism, 2007, pp. 50-53, 76-80).

Maar ook de kwaliteit van peer review is niet onomstreden. Mogelijke problemen in het geven en ontvangen van peer review zijn de mate van subjectiviteit, partijdigheid en kwetsbaarheid in de beoordeling, communicatie en ontvangst van de peer review (Finn & Garner, 2011; Van Note Chism, 2007, pp. 20-22). Als voorbeeld van kwetsbaarheid kan worden genoemd dat het voor een arts-assistent in de opleiding bedreigend, vervelend en moeilijk kan zijn om een oordeel te moeten vellen over de kwaliteit van het onderwijs van de hoogleraar van de vakgroep, die op zijn of haar beurt de

summatieve opleidingsinspanningen van de arts-assistent moet beoordelen. Veiligheid, vertrouwen, respect en gelijkheid zijn dan ook essentieel voor het goed kunnen uitvoeren van een peer review (Alabi & Weare, 2014; Siddiqui, Jonas-Dwyer, & Carr, 2007; Van Note Chism, 2007, pp. 21, 33, 49, 100-101).

Met het borgen van veiligheid, vertrouwen, respect en gelijkheid is het probleem van subjectiviteit op de uitkomst en kwaliteit van de peer review echter nog niet opgelost. Verschillen in visie en weging van criteria in de beoordeling van onderwijs tussen de docent en de peer review gevende collega, kunnen het geven en ontvangen van een peer review beïnvloeden (Alabi & Weare, 2014; Newman et al., 2009). Het ontwikkelen van lokaal ontworpen en lokaal geïnterpreteerde criteria voor het geven en ontvangen van peer review wordt dan ook aanbevolen (Van Note Chism, 2007, pp. 62-64).

Ondanks de aanbevelingen vanuit de wetenschap voor de introductie en het gebruik van peer review in onderwijs (Berk, 2013; Gusic et al., 2014; Van Note Chism, 2007) zijn er weinig vragenlijsten te vinden voor het geven en ontvangen van een peer review van de kwaliteit van hoorcolleges in academisch-medisch onderwijs (Newman et al., 2009). In de voorstudie voor het onderzoek, waarbij meer dan 350 mogelijk relevante wetenschappelijke artikelen werden geraadpleegd van de afgelopen 10 jaar, kon uitsluitend de door Newman et al. (2009) ontwikkelde vragenlijst gevonden worden voor het geven van peer review van hoorcolleges binnen academisch-medisch onderwijs (zie Tabel 1). Dit instrument is echter slechts op kleine schaal ($N$ = 31) onderzocht op betrouwbaarheid, en dit onderzoek gebeurde op basis van het scoren van de kwaliteit van onderwijs aan de hand van video-opnames (Newman et al., 2009). Uit dit onderzoek naar betrouwbaarheid bleken de interpersoonlijke scores over hetzelfde onderwijs op vier van de elf vragen uit de lijst weinig met elkaar te correleren (Newman et al., 2009). De vragenlijst bevat ook vijf meervoudig gestelde vragen waardoor de betrouwbaarheid in de beantwoording negatief kan worden beïnvloed. De vragenlijst richt zich daarnaast vooral op de fysieke prestaties van de docent op het moment van lesgeven en minder op de relatie tussen het verzorgde onderwijs en de geldende leertheorieën. Het door Newman et al. (2009) ontwikkelde instrument onderzoekt bijvoorbeeld niet de relatie tussen het gegeven onderwijs en de aanwezigheid van de benodigde voorkennis van de student of de kwaliteit van de leeromgeving als factoren die van invloed kunnen zijn op de kwaliteit van leren (Driscoll, 2014). Aanvullend punt van aandacht bij het gebruik van deze vragenlijst in Nederland is het feit dat het Engelstalige instrument ontwikkeld is in en voor de Amerikaanse medische context, die andere eisen en behoeften kan hebben dan de Nederlandse medische onderwijssituatie (Molenaar et al., 2009).

Om het geven van peer review van de kwaliteit van hoorcolleges binnen de academisch-medische opleiding van ons UMC te ondersteunen was behoefte aan de ontwikkeling van lokaal ontwikkelde en lokaal geïnterpreteerde criteria (Van Note Chism, 2007, pp. 62-64). Een vragenlijst die deze criteria verwoordt, zou een bijdrage kunnen leveren aan de verdere professionalisering van het geven en ontvangen van peer review van hoorcolleges binnen ons UMC. Met deze vragenlijst zou de

kwaliteit van hoorcolleges binnen het medisch onderwijs van ons UMC breder kunnen worden beoordeeld dan met het gebruik van uitsluitend studentenevaluaties (Gusic et al., 2014; Van Note Chism, 2007). Om deze redenen had de groep van PE's binnen ons UMC van de Raad van Bestuur toestemming gekregen onderzoek te verrichten naar de constructie van een nieuwe vragenlijst ter ondersteuning van het geven en ontvangen van een vrijwillige, formatieve peer review van de kwaliteit van hoorcolleges binnen de academisch-medische opleiding. Op verzoek van de groep van docenten staat het toekomstig gebruik van de vragenlijst in ieder geval de eerste jaren nog los van de binnen ons UMC aanwezige kwaliteitskaders om een onafhankelijke ontwikkeling van de vragenlijst mogelijk te maken en docenten te laten wennen aan het ontvangen en verzorgen van een regelmatige peer review op de kwaliteit van de hoorcolleges. De uiteindelijke vragenlijst zal de eerste jaren daarmee nog geen formele status krijgen binnen de kwaliteitskaders van ons UMC. Deze scriptie beschrijft de eerste fase van de totstandkoming van deze vragenlijst, waarin de vragenlijst ontwikkeld wordt, en onderzocht zal worden op interne betrouwbaarheid en factoren.

Tabel 1
*De vragen uit de vragenlijst van Newman et al. (2009)*

| Vragen |
|---|
| Clearly states goals to the talk |
| Communicates or demonstrates importance of the lecture's topic(s) |
| Presents material in a clear, organized fashion |
| Shows enthusiasm for the topic |
| Demonstrates command of the subject matter |
| Explains and summarizes key concepts |
| Encourages appropriate audience interaction |
| Monitors audience's understanding of material and responds accordingly |
| Audio and/or visual aids reinforce the content effectively |
| Voice is clear and audio-visuals are audible/legible |
| Provides a conclusion to the talk |

**Vraagstellingen bij het Onderzoek**

De centrale vraagstelling voor het onderzoek was: Hoe kan het geven van een peer review van de kwaliteit van een hoorcollege in onze academisch-medische opleiding worden ondersteund met behulp van een korte vragenlijst van maximaal 10 items? Deelvragen bij dit onderzoek waren: (a) Wat zijn de verschillende factoren die uit deze vragenlijst kunnen worden afgeleid? (b) Wat is de interne betrouwbaarheid van deze factoren binnen deze vragenlijst? en (c) Hoe verhouden de uitkomsten van deze vragen zich tot de bestaande literatuur?

## Methode

### Ontwerp

Het onderzoek omvatte vier stadia, ieder met afzonderlijke methoden: (a) uitgebreid exploratief kwalitatief literatuuronderzoek naar wetenschappelijk onderbouwde beoordelingskaders, kwaliteitskaders en vragenlijsten op het terrein van evaluatie van de kwaliteit en effectiviteit van (medisch) onderwijs; (b) een iteratief proces[1] waarbij, aan de hand van het eerdere literatuuronderzoek, een eigen initiële nieuwe vragenlijst is ontwikkeld op basis van de meest voorkomende vragen uit de literatuur; (c) het testen van de vragenlijst, door het laten beoordelen van hoorcolleges, met behulp van de nieuw ontwikkelde vragenlijst, door collega-docenten binnen ons UMC; en (d) het, met behulp van kwantitatief correlatieonderzoek (Creswell, 2014), analyseren van de gegevens, onderzoeken van mogelijke factoren en berekenen van de betrouwbaarheid van de factoren met behulp van Cronbach's α. Omdat de vragenlijst primair bedoeld was voor gebruik binnen de populatie waarbinnen ook de validatie plaatsvond werden de factoren onderzocht met behulp van Principle Axis Factoring (Field, 2013). Het gehele onderzoek kan worden aangemerkt als *design-based onderzoek*, omdat er sprake is van systematisch onderzoek naar een bestaand probleem uit de eigen praktijksituatie, gevolgd door de ontwikkeling en het valideren van, en het reflecteren op, een onderwijskundig instrument dat tot doel heeft dit probleem op te lossen en de praktijk van onderwijs te verbeteren (Plomp, 2007).

De studie is geheel volgens planning uitgevoerd, op aanvraag en onder begeleiding van Prof. Dr. M.P. Schijven (PE, chirurg en MSc Public Health Education and promotion), Dr. E.J.M. Nieveen van Dijkum (PE, chirurg) en Drs. J.A. Baane (senior onderwijskundige), en met goedkeuring van de Raad van Bestuur en de Functionaris Gegevensbescherming van het UMC. Het onderzoek werd uitgevoerd in de periode december 2017-juni 2019. Met toestemming van de examencommissie en de cursuscoördinator voor de masterthesis van de Open Universiteit mocht ik dit onderzoek gebruiken voor mijn thesis voor de master Onderwijswetenschappen. Deze scriptie is geheel door mijzelf geschreven, zonder hulp van de opdrachtgevers en begeleiders. Dat de werkzaamheden tijdens het onderzoek daadwerkelijk door mijzelf zijn uitgevoerd wordt ondersteund in de ondertekende en gespecificeerde verklaring die in het bezit is van de Open Universiteit.

[1] Dit iteratief proces hield in dat een vast onderzoeksteam van professionals, bestaande uit twee ervaren PE's, een onderwijskundige en een ervaren docent/onderwijskundige in opleiding, meerdere keren, doch minimaal eens per maand, bij elkaar zijn gekomen gedurende de gehele looptijd van het onderzoek (december 2017-juni 2019). Tussen de bijeenkomsten hielden de groepsleden, mondeling en via e-mail, contact met elkaar en met andere docenten en onderwijskundigen binnen ons UMC en heeft de student, als uitvoerder van het onderzoek, de opmerkingen verwerkt in nieuwe versies van vragen en vragenlijsten. Pas wanneer consensus was ontstaan over de constructie van de vragen en de samenstelling van de vragenlijsten werd deze geaccordeerd voor verder gebruik.

**Participanten**

Het literatuuronderzoek is uitsluitend door mijzelf als onderzoeker uitgevoerd. Deelnemers aan het iteratief proces van de totstandkoming van de initiële vragenlijst zijn twee PE's, een ervaren academisch docent/onderwijskundige in opleiding (student van de Open Universiteit) en drie ervaren onderwijskundigen van ons UMC, waarvan er een deel uitmaakte van het vaste onderzoeksteam. Gedurende de looptijd van december 2017-juni 2019 werd de onderzoeker begeleid door het vaste onderzoeksteam van twee ervaren PE's en een ervaren onderwijskundige van het UMC. Deelnemers aan de valideringsfase van de vragenlijst waren docenten uit de medische opleiding van ons UMC (*N* = 37, leeftijd 29-62 jaar, 40% man, 60% vrouw) die, op geheel vrijwillige basis en na voorlichting over het doel, methode en anonimiteit van onderzoek, zich wilden laten beoordelen door collega's. Zevenendertig docenten, die 37 unieke hoorcolleges verzorgden, werden binnen de periode van augustus-oktober 2018 gekozen op basis van de planning van de hoorcolleges, waarbij een minimale lestijd van 20 minuten en een minimaal aantal van vijf aanwezige collega-docenten werden aangehouden als criteria. De collega's waren eveneens docenten uit de medische opleiding van ons UMC (*N* = 195, leeftijd 25-66 jaar, 40% man, 60% vrouw) die, op geheel vrijwillige basis en na voorlichting over het doel, methode en anonimiteit van onderzoek, wilden bijdragen aan de validatie door het invullen van de beoordeling. Zowel de docenten als de collega-docenten kwamen uit 17 verschillende vakgroepen binnen ons UMC (11 klinische en 6 preklinische vakgroepen). De collega-docenten die de beoordeling verrichten waren volgens planning aanwezig tijdens het hoorcollege. Als docenten of collega's werden aangemerkt: alle (externe) stafleden, professionals en arts-assistenten die hoorcolleges binnen de academisch-medische opleiding verzorgden of daarbij aanwezig waren. Coassistenten en studenten werden uitgesloten voor de beoordeling van het onderwijs omdat hun ontwikkelingsniveaus op het gebied van onderwijs niet gelijk werden geacht aan die van de docenten en zij hun mening over het onderwijs al konden geven in de studentenevaluaties.

**Materialen**

Gebruik is gemaakt van bekende databases voor wetenschappelijke literatuur als ERIC (OVID), PsycINFO (OVID), Web of Science en PubMed. Daarnaast is op internet en in de bibliotheek van het UMC gezocht naar in de literatuur beschreven beoordelingskaders en vragenlijsten. Op basis van de gevonden literatuur en een iteratief proces binnen het vaste onderzoeksteam, waarbij twee extra onderwijskundigen van ons UMC werden betrokken (zie Procedure), is een vragenlijst van 21 items ontwikkeld voor de ondersteuning van peer review van academisch-medische hoorcolleges (zie ook Procedure en Figuur 1). Aan deze vragenlijst is, in consensus binnen het vaste onderzoeksteam, een 5-punts-Likert-scoringsschaal toegevoegd met vijf gerelateerde scores (--, -, 0, + en ++). Deze initiële vragenlijst werd voorzien van drie negatief gestelde vragen om de beoordelaar alert te houden (vraag 6, 9 en 18) en een open ruimte waarin de beoordelaar eventuele aanvullende opmerkingen, uitleg of suggesties kon vermelden voor de docent. Deze initiële vragenlijst is gebruikt in het validatieproces.

Beoordelingskaders en vragenlijsten zijn verzameld in een Excel-bestand. IBM SPSS versie 22 is gebruikt voor de analyse van de resultaten van het validatieproces, de berekening van de betrouwbaarheid en de berekening van de factoren. Endnote 8 is gebruikt voor het bijhouden van gevonden literatuur.

**Procedure**

De procedure van het onderzoek omvatte de volgende stadia:

1. December 2017-april 2018: De onderzoeker verrichte zelf een uitgebreide literatuurstudie in ERIC (OVID), PsycINFO (OVID) en Pubmed, naar bestaande, op wetenschappelijke basis ontwikkelde vragenlijsten op het gebied van evaluatie van de kwaliteit en effectiviteit van (medisch) onderwijs. Hierbij zijn, onder begeleiding van een klinische bibliothecaresse, de zoektermen gebruikt zoals die vermeld staan in Tabel 2. Veelvuldige aanpassingen van deze zoektermen leidden niet tot betere resultaten. Om een verdere selectie van artikelen mogelijk te maken werden in consensus binnen het vaste onderzoeksteam criteria opgesteld voor de selectie van artikelen voor verder gebruik (zie Tabel 3). Van alle gevonden artikelen werd de samenvatting beoordeeld op deze criteria. Gezocht werd naar, met goede wetenschappelijke methodiek onderbouwde, studentenevaluaties, vragenlijsten voor peer review, en professionele beoordeling- en kwaliteitskaders. Onderzoeken naar het meten van kwaliteit en effectiviteit van onderwijs in coschappen of medische specialisatie-opleidingen werden uitsluitend meegenomen wanneer hoorcolleges waren onderzocht. Met behulp van backward searches werd gezocht naar aanvullende relevante wetenschappelijke literatuur in ERIC (OVID), PsycINFO (OVID), Web of Science, Pubmed, internet en de bibliotheek van het UMC. Aanvullende literatuur over de (totstandkoming van de) vragenlijsten uit de BKO en UvA-Q[2] werd verkregen van bronnen binnen ons UMC of de overkoepelende universiteit;
2. Maart 2018-mei 2018: Om de later in de literatuur te vinden vragen te kunnen indelen naar categorieën waarover de vragen informatie verzamelden, is gestart met de ontwikkeling van een categorieënoverzicht voor de vragen. Met de constructie van dit categorieënoverzicht wilden wij een latere selectie van vragen die van toepassing waren op het moment van instructie vergemakkelijken.

   Door de onderzoeker zelf werden alle in de literatuur gevonden kaders op het gebied van de beoordeling van de kwaliteit van medisch onderwijs beoordeeld op hun wetenschappelijke onderbouwing (theoretische onderbouwing, methodiek, empirische toetsing). Vervolgens werd onderzoek verricht naar de doelen waarvoor de beoordelingskaders waren ontworpen en de verschillen en overeenkomsten tussen de benamingen van de categorieën van de beoordelingskaders. Alle categorieën uit de beoordelingskaders werden zo nodig vertaald naar

[2] De Uva-Q is een basisvragenlijst die binnen de overkoepelende universiteit van ons UMC wordt gebruikt. Uit de UvA-Q kunnen vragen voor studentenevaluaties worden geselecteerd.

het Nederlands. Deze categorieën werden door de onderzoeker geordend in een Excel-bestand. Er werd geen vergelijkend onderzoek uitgevoerd naar de verschillen in inhoud binnen de categorieën. In een iteratief proces binnen het vaste onderzoeksteam werd in een drietal sessies in consensus, op basis van de meest voorkomende categorieën in bestaande wetenschappelijk goed onderbouwde beoordelingskaders, een categorieënoverzicht ontwikkeld voor rubricering van de in de literatuur te vinden vragen. Elkaar aanvullende categorieën werden gekozen voor het vormen van dit categorieënoverzicht. In de constructie van dit categorieënoverzicht werd geen rekening gehouden met peer review of het moment van beoordeling;

3. Maart 2018-mei 2018: De gevonden vragenlijsten werden door de onderzoeker beoordeeld op wetenschappelijke onderbouwing. Alle gevonden vragen werden indien nodig vertaald naar het Nederlands. De vragen werden in een Excel-bestand gerubriceerd naar de categorieën van het eerder ontwikkelde categorieënoverzicht. Binnen deze categorieën werden de vragen gerangschikt naar onderwerp om een goed inzicht te kunnen krijgen van de gehele verzameling vragen. Vervolgens werd door de onderzoeker aangegeven welke vraag door wie kon worden beantwoord (student, collega-docent, onderwijskundige of de docent zelf);

Tabel 2
*Zoektermen voor het literatuuronderzoek, op basis van de voorstudie*

| Zoektermen voor het literatuuronderzoek |
|---|
| (medical education AND (peer review OR peer assessment) AND (teacher OR lecturer) AND limit to English and yr="2007 -Current") NOT (student(ti,ab) OR resident (ti,ab) |
| ((evaluation criteria OR lesson observation criteria OR standard* OR best practice* OR educational principles OR accomplished teach* OR gifted teach* OR skillful teach* OR talented teach* OR enhanced teach* OR efficiency of teach* OR competence of teach* OR performance of teach* OR powerfulness of teach* OR effectiveness of teach* OR quality of teach*) adj6 teach*).ti,ab,id. AND (professional education OR medical education OR teacher education OR higher education OR graduate study OR postdoctoral education OR college faculty OR faculty development) AND (teach* effect OR teach* quality) AND (limit to English and yr="2007 -Current") |
| (docent OR doceren OR leraar OR opleider OR opleiden).ti,ab,hw,id. AND (beoordeling OR waardering OR kwaliteit OR talent OR capaciteit OR niveau).ti,ab,hw,id. AND (limit to full text and yr="2007 -Current") |
| (higher education OR medical education) AND (educational performance OR educator evaluation OR performance evaluation OR teaching effectiveness OR teaching quality).ti,ab,id. AND (limit to English and yr="2007 -Current") |
| (docent OR doceren OR opleiden OR opleider OR leraar).ti,ab,hw,id. AND (beoordeling OR waardering OR kwaliteit OR talent OR capaciteit OR niveau).ti,ab,hw,id. AND (limit to full text and yr="2007 -Current") |
| (Medical Education[Majr] OR Medical Faculty[Majr]) AND (educator performance[tiab] OR educator evaluation[tiab] OR evaluating educator*[tiab] OR teaching effectiveness[tiab]) AND ("2007/01/01:2017/12/31"[Date – Publication]) AND (English[Language]). |

Tabel 3
*Criteria voor inclusie van de gevonden literatuur*

| Criteria voor inclusie van gevonden literatuur |
| --- |
| Beschrijft het artikel de ontwikkeling of evaluatie van onderzoeksmiddelen naar de beoordeling van de kwaliteit van (medisch) onderwijs? |
| Is de in het artikel gebruikte onderzoeksmethode wetenschappelijk valide en zijn de uitkomsten van het onderzoek betrouwbaar? |
| Is het artikel in volledige tekst beschikbaar binnen de databases of via internet? |
| Betreft het artikel onderzoek naar de beoordeling van kwaliteit van hoorcolleges? Studies naar de kwaliteit van individueel of groepsonderwijs binnen coschappen en arts-assistentschappen werden uitgesloten omdat deze onderwijsvormen een andere opzet hebben dan hoorcolleges. |

4. Maart 2018-juni 2018: Na ordening van de in de literatuur gevonden 701 vragen binnen het ontworpen categorieënoverzicht door de onderzoeker, werd de indeling van de vragen door het vaste onderzoeksteam in consensus vastgesteld. In consensus werden 25 vragen geselecteerd die door een collega-docent beoordeeld zouden kunnen worden op het moment van instructie. Vragen die uitsluitend door een onderwijskundige of student beantwoord zouden kunnen worden werden uitgesloten voor verder gebruik;
5. Mei 2018-juli 2018: In consensus binnen het vaste onderzoeksteam, en met ondersteuning van twee andere onderwijskundigen van ons UMC, werd in drie iteratieve sessies een initiële vragenlijst van 21 vragen samengesteld voor de ondersteuning van peer review van medische hoorcolleges binnen ons UMC. Van de eerder geselecteerde 25 vragen werden 4 vragen niet verder meegenomen in het proces omdat zij een dubbele vraagstelling bevatten of niet van toepassing werden geacht binnen ons UMC[3]. In consensus werd een 5-punts-Likert-schaal met criteria toegevoegd aan de vragen (--, -, 0, +, ++). De optie niet van toepassing werd toegevoegd aan de vijf gekozen criteria en een extra open vraag werd toegevoegd om de beoordelaars de mogelijkheid te geven aanvullende kwalitatieve feedback te geven op de gegeven beoordelingen. Daarnaast werd besloten om drie vragen negatief geformuleerd te stellen om de gebruiker van de vragenlijst alert te houden tijdens het invullen van de vragenlijst (zie ook Figuur 1). *Psychology of Learning for Instruction* (Driscoll, 2014) werd gebruikt als leidraad om, waar nodig, consensus te krijgen over de motivatie van de samenstelling van de initiële vragenlijst;

[3] Voorbeeld van een dergelijke vraag is bijvoorbeeld: “De docent kan uitstekend met de audiovisuele middelen omgaan”. Omdat docenten tijdens het geven van hoorcolleges worden ondersteund door vakmensen van de audiovisuele dienst werd de beantwoording van deze vraag als onvoldoende relevant beschouwd voor het gebruik binnen peer review.

| Datum: Docent: | | mee oneens / eens | | | | | |
|---|---|---|---|---|---|---|---|
| | **Stelling:** | -- | - | 0 | + | ++ | nvt |
| 1 | Bij aanvang is de docent helder over de inhoud van het onderwijs | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 2 | Bij aanvang is de docent helder over de leerdoelen van het onderwijs | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 3 | Bij aanvang is de docent helder over de wijze van toetsing van het onderwijs | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 4 | Om een overgang te maken naar de nieuw te leren stof activeert de docent waar nodig eerst de benodigde voorkennis | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 5 | De docent spreekt verstaanbaar | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 6 | De docent laat achterwege om de hoofd- en bijzaken aan te geven | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 7 | Er worden door de docent verschillende visies op de problematiek aangereikt | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 8 | Met goede verhelderende voorbeelden uit de praktijk illustreert de docent de essenties van de leerstof | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 9 | Ik heb moeite het onderwijs goed te begrijpen | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 10 | De docent onderzoekt regelmatig of studenten het onderwijs begrijpen | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 11 | Om problemen in het kunnen begrijpen van het onderwijs op te lossen maakt de docent indien nodig direct aanpassingen in het onderwijs | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 12 | De docent creëert een veilige leeromgeving | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 13 | De docent stimuleert actieve deelname aan het onderwijs | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 14 | De docent houdt studenten actief betrokken bij het onderwijs | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 15 | De docent stimuleert samenwerkend leren en onderling overleg over de lesstof | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 16 | Het onderwijs is gebaseerd op een kritische beschouwing van hedendaagse wetenschappelijke literatuur, onderzoek en best practises | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 17 | Studenten worden door de docent tot zelfstandig nadenken gestimuleerd | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 18 | De docent straalt weinig enthousiasme uit voor het vakgebied | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 19 | Naar alle individuele studenten, patiënten en collega's is de docent respectvol | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 20 | Leerdoelen, leervormen en toetsing vormen een samenhangend geheel | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 21 | Met de beschikbare tijd komt de docent goed uit | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| Aanvullende opmerkingen en/of voorbeelden bij de beantwoording van de vragen: | | | | | | | |

*Figuur 1.* De initiële vragenlijst zoals die in het onderzoek ontworpen, getest en geanalyseerd is.

6. In de periode mei 2018-juli 2018 heeft de onderzoeker de staven en verantwoordelijken voor onderwijs binnen 12 klinische en 11 preklinische vakgroepen mondeling en via email benaderd om binnen de vakgroepen toestemming te krijgen voor deelname van docenten en collega-docenten aan de validatiefase van de initiële vragenlijst;
7. Na toestemming van de vakgroepen werden in de periode augustus 2018-oktober 2018 docenten die gepland stonden voor het geven van onderwijs ($N$ = 37, leeftijd 29-62 jaar, 40% mannen en 60% vrouwen) persoonlijk en vertrouwelijk door de onderzoeker per mail

benaderd met het verzoek hun onderwijs te laten beoordelen door collega's met behulp van de vragenlijst. Een week voor het geplande onderwijs vond per mail een eerste contact tussen de onderzoeker en de docent plaats, waarin uitleg werd gegeven over de opzet, doelen en vertrouwelijkheid van de deelname. Daarnaast werd in deze mail een eerste, voorlopig akkoord voor deelname gevraagd. Wanneer de docent open stond voor deelname werd de docent door de onderzoeker in een aanvullend persoonlijk gesprek verder ingelicht over het onderzoek en in de gelegenheid gesteld vragen te stellen. Deelname voor alle docenten was vrijwillig en vertrouwelijk en intrekken van deelname was te allen tijde mogelijk. De informatie over de planning van het onderwijs verkreeg de onderzoeker uit de systemen van het UMC of uit gedeelde en open informatie van deelnemende vakgroepen. Alleen onderwijsmomenten met een geplande tijdspanne van minstens 20 minuten en met een aanwezigheid van minimaal vijf collega-docenten werden geselecteerd voor het validatieproces. Het minimale aantal van vijf collega-docenten werd aangehouden om de vertrouwelijkheid van de invullers te kunnen garanderen;

8. Na toestemming van de docent voor deelname werd direct voorafgaand aan het moment van onderwijs de vragenlijst aan de aanwezige collega-docenten uitgereikt met het verzoek deze achteraf in te vullen. Collega-docenten ($N$ = 195, leeftijd 25-66 jaar, 40% mannen en 60% vrouwen) werden door de onderzoeker ingelicht over doelen, opzet en vertrouwelijkheid van de deelname aan het validatieproces. Geen verdere informatie werd verstrekt over de inhoud en mening van vragen, om het validatieproces niet te verstoren. De vragenlijst werd direct na het moment van instructie ingevuld en door de onderzoeker persoonlijk verzameld. Deelname aan het validatieproces door de collega-docenten was vrijwillig en vertrouwelijk en intrekken van deelname was te allen tijde mogelijk. Deelnemende docenten die hun onderwijs lieten beoordelen ontvingen van de onderzoeker eenmalig in vertrouwen een geanonimiseerde samenvatting van de beoordelingen van hun collega-docenten per mail. Uitgaande van de stelling van Field (2013, p. 683) werd een beoogd aantal ingevulde vragenlijsten aangehouden van minimaal 15 maal het aantal elementen van de vragenlijst, met een minimum van 300 vragenlijsten;
9. November 2018-maart 2019: Analyse door de onderzoeker zelf (zie Analyse).
10. Maart 2019: Omdat er door de opdrachtgevers aan de opdracht werd vastgehouden de vragenlijst terug te brengen naar een maximaal aantal van 10 items, werd, tegen het advies van de onderzoeker in, de resterende vragenlijst voorgelegd aan een buiten het vaste onderzoeksteam vallende discussiegroep van onderwijskundigen en ervaren academische docenten (leeftijd 40-59 jaar, twee onderwijskundigen, twee docenten, twee mannen, twee vrouwen). Deze discussiegroep was niet op de hoogte van het voortraject van de ontwikkeling van de vragenlijst of de eerder gemaakte indeling in categorieën. Deze discussiegroep kreeg van de onderzoeker de vraag voorgelegd hoe de vragenlijst van 15 naar 10 items kon worden

teruggebracht zonder veel in waarde te verminderen. Daarnaast werd deze discussiegroep gevraagd de mogelijke constructen van de te berekenen factoren in de uiteindelijke vragenlijst te benoemen. De resultaten van deze discussie werden voorgelegd aan het vaste onderzoeksteam om in consensus te worden geaccordeerd. *Educational Research: Planning, Conducting and Evaluating Quantitative and Qualitative Research* (Creswell, 2014) en *Discovering Statistics Using SPSS* (Field, 2013) werden gebruikt als leidraad in de uitvoering van de analyse. *Psychology of Learning for Instruction* (Driscoll, 2014) werd gebruikt als referentie op het gebied van leertheorieën.

**Analyse**

In de periode november 2018-februari 2019 werden alle ingevulde vragenlijsten geanonimiseerd ingevoerd in SPSS, versie 22. Iedere unieke docent werd met een opeenvolgend nummer binnen SPSS onderscheiden. Scores op de 5-punts-Likertschaal werden omgezet naar de waarden -2 (--), -1 (-), 0 (0), 1 (+) en 2 (++), met 9 als waarde voor de score niet van toepassing en 99 als waarde voor een ontbrekend antwoord. Scores op de negatief gestelde vragen (vraag 6, 9 en 18) werden hierbij omgekeerd verwerkt, rekening houdend met het feit dat hierbij een negatief antwoord juist een positieve beoordeling inhield en vice versa.

Kwantitatieve analyse vond plaats van de dataset, waarbij verdeling, *skewness* (scheefheid van de verdeling), *kurtosis* (spitsheid van de verdeling), onderlinge correlaties en aannamen voor een normaalverdeling werden onderzocht (Field, 2013). Paarsgewijze exclusie werd toegepast bij het ontbreken van data om voldoende steekproefgrootte te kunnen borgen. Na bevestiging van de aannamen voor een normaalverdeling werd een *Principle Axis Factoring* (*PAF*) uitgevoerd met *Direct Oblimin*, *Kaiser Normalisatie*, δ 0, 25 iteraties en paarsgewijze exclusie (Field, 2013). De factoranalyse werd uitgevoerd om te controleren of er een of meerdere mogelijke factoren berekend konden worden uit de gegevens van de vragenlijst, omdat de aanwezigheid van factoren van invloed kan zijn op de interne betrouwbaarheid van de vragenlijst (Field, 2013, p. 709). Daarnaast wilden wij weten of uit de vragenlijst constructen geformuleerd konden worden die met de vragenlijst zouden kunnen worden gemeten. Omdat de (analyse van de) vragenlijst primair bedoeld was voor conclusies en gebruik binnen de populatie waarbinnen ook de validatie plaatsvond werden de factoren onderzocht met behulp van Principle Axis Factoring (Field, 2013). Als criteria voor de bepaling van het aantal factoren werden aangehouden: (a) het buigpunt in de *Screeplot*; (b) eigenwaarden met een minimum van 1,0; (c) een minimale aanvullende verklaarde variantie van 5% per factor; (d) een afbreekpunt voor factorladingen van ,3 (Field, 2013, p. 681); (e) items die binnen dezelfde factor aanwezig zijn meten mogelijk een soortgelijk construct; (f) variabelen die binnen verschillende factoren aanwezig zijn meten mogelijk verschillende constructen; en (g) variabelen laden niet op verschillende factoren met waarden groter dan ,3. Op basis van de analyse werden onbetrouwbare en onvoldoende onderscheidende vragen uit de vragenlijst verwijderd en werd de hierboven beschreven analyse

herhaald op de resterende vragen. In het proces van de aanpassing van de vragenlijst tot de gewenste grootte van 10 vragen werden opnieuw alle individuele factoren berekend en werden in consensus constructen bedacht die gereflecteerd konden worden door de gevonden factoren. Het berekenen van de *Inter-rater reliability* (IRR) was, ondanks de hulp van een docent methodologie van de Open Universiteit, helaas niet mogelijk met de samengestelde dataset omdat er sprake was van een *shared between level construct* in de opzet van het onderzoek.

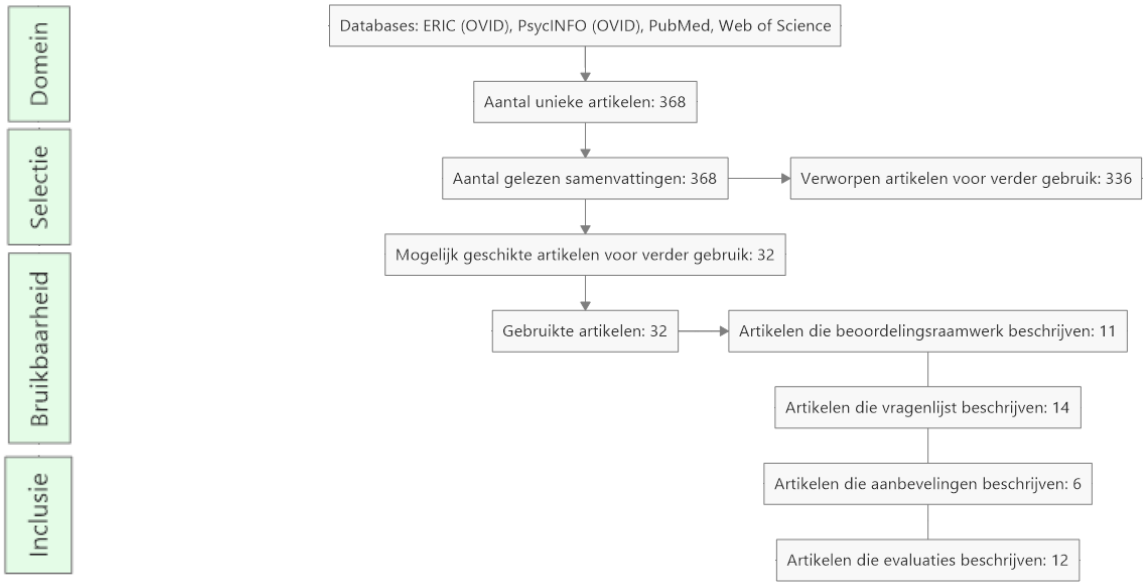
## Resultaten

De Engelstalige zoekstrategieën leverden in totaal 368 mogelijk relevante artikelen op (zie Figuur 2). Geen van de Nederlandstalige zoekstrategieën leverde enig resultaat binnen de gekozen databases op. Op basis van de gestelde criteria werden uit de databases 32 artikelen geselecteerd voor verder onderzoek. Twee aanvullende artikelen werden verkregen via contacten binnen de universiteiten van Amsterdam: (a) Bestuursstaf, 2012; en (b) Van de Wiel, de Jong, Mulder, en Schlusmans (Eds.), 2016. De UvA-Q en De Epicuris-vragenlijst[4] waren niet beschikbaar op internet en werden verkregen via contacten binnen de overkoepelende universiteit en ons UMC.

In het totaal van 34 verkregen artikelen werden door 11 artikelen in totaal acht verschillende beoordelingskaders beschreven voor de beoordeling van de kwaliteit van onderwijs (Academy of Medical Educators, 2014; Baldwin et al., 2011; Bestuursstaf, 2012; Frank, Snell, & Sherbino, 2015; Gusic et al., 2014; Mintz, Southern, Ghali, & Ma, 2015; Molenaar et al., 2009; Owolabi, 2014; Pettit et al., 2015; Srinivasan et al., 2011; Van de Wiel et al., 2016). De verschillende beoordelingskaders omvatten in totaal 29 verschillende beoordelingscategorieën voor de beoordeling van de kwaliteit van onderwijs (zie Tabel 4). Drie aanvullende artikelen beschreven geen compleet beoordelingskader, maar criteria of vragenlijsten waarmee de kwaliteit van onderwijs op kleinere schaal kon worden onderzocht (Newman et al., 2009; Simendinger et al., 2017; Valiee, S., Moridi, G., Khaledi, S., & Garibi, F., 2015). De UVA-Q en de Epicuris-vragenlijst, gevalideerde vragenlijsten verkregen binnen ons UMC en de overkoepelende universiteit, werden toegevoegd aan de gevonden artikelen.

De gevonden beoordelingskaders waren allen goed wetenschappelijk onderbouwd door uitgebreide en langdurige iteratieve ontwerpprocessen en empirisch onderzoek. De beoordelingskaders toonden grote verschillen onderling. Op basis van de prevalentie en inhoud werden in consensus uiteindelijk 10 elkaar aanvullende categorieën gekozen voor het samenstellen van ons eigen categorieënoverzicht. Met de samenstelling van het eigen categorieënoverzicht werd geen rekening gehouden met het moment van beoordeling of de beoordeling door middel van peer review. Het categorieënoverzicht was uitsluitend bedoeld als hulpmiddel voor het kunnen indelen van de gevonden vragen en criteria op het gebied van de beoordeling van de kwaliteit van onderwijs in het algemeen (zie Tabel 5). De categorieën van het ontwikkelde categorieënoverzicht correspondeerden het meest

[4] De Epicuris-vragenlijst is een gevalideerde vragenlijst die binnen ons UMC gebruikt wordt ter evaluatie van het onderwijs in de bachelorfase




*Figuur 2.* Prismadiagram voor de literatuurstudie naar beoordelingskaders en vragenlijsten. Omdat artikelen met een beoordelingskader ook vragen en criteria uit dit beoordelingskader beschrijven staan deze dubbel vermeld in de inclusie en komt het totaal aantal artikelen in de inclusie niet overeen met het aantal gebruikte artikelen.

met de categorieën uit de beoordelingskaders zoals die omschreven zijn door: (a) de AAMC (Baldwin et al., 2011; Gusic et al., 2014); (b) Molenaar et al. (2009); en (c) de Academy of Medical Educators (2014).

Binnen de geselecteerde literatuur werden 14 artikelen met criteria of vragenlijsten gevonden (Academy of Medical Educators, 2014; Baldwin et al., 2011; Bestuursstaf, 2012; Frank et al., 2015; Gusic et al., 2014; Mintz et al., 2015; Molenaar et al., 2009; Newman et al., 2009; Owolabi, 2014; Pettit et al., 2015; Simendinger et al., 2017; Srinivasan et al., 2011; Valiee et al., 2015; Van de Wiel et al., 2016). Elf van deze artikelen waren ook gebruikt voor de samenstelling van het eigen categorieënoverzicht. Alle artikelen waren wetenschappelijk goed onderbouwd vanuit een langdurig iteratief proces of onderzoek. De gevalideerde vragenlijsten van de UVA-Q en Epicuris werden aan de lijst toegevoegd. De vragenlijsten omvatten in totaal 701 criteria of vragen op het gebied van onderzoek naar de kwaliteit van onderwijs. Na rubricering konden uit het totaal aantal criteria en vragen 75 unieke vragen worden geabstraheerd. Vijfentwintig van deze vragen betroffen het moment van instructie die, naar de mening van het vaste onderzoeksteam, door collega-docenten beantwoord zouden kunnen worden. Deze vragen waren afkomstig uit de categorieën leerklimaat, prestaties tijdens instructie en studentgerichtheid. In de overige categorieën werden geen vragen gevonden die van toepassing waren op het doceren op het moment van instructie. In een iteratief proces werden 21 vragen gekozen voor verder gebruik in het validatieproces. Vier van de eerder geselecteerde 25 vragen werden buitengesloten voor verder gebruik omdat zij een dubbele vraagstelling bevatten of niet van toepassing werden geacht op de situatie binnen ons UMC. Vragen 6, 9 en 18 werden gekozen voor een negatieve formulering van de vraagstelling (zie Methode en Figuur 1).

Tabel 4
*Beoordelingskaders en hun categorieën*

| | Academy of Medical Educators, 2014 | BKO (Bestuursstaf, 2012; Van de Wiel, 2016) | CanMeds (Frank et al., 2015) | AAMC (Baldwin et al., 2011; Gusic et al., 2014) | SFDPQ (Mintz et al., 2015; Owolabi, 2014) | Molenaar et al., 2009 | Pettit et al., 2015 | Srinivasan et al., 2011 |
|---|---|---|---|---|---|---|---|---|
| Adequate leermethodieken | | | | x | | | | |
| Begeleiden van studenten | | x | | | | | | |
| Bevorderen van begrip en onthouden | | | | | x | | | |
| Bevorderen van zelfgestuurd leren | | | | | x | | | |
| Bijdrage aan onderzoek en verspreiding van de resultaten ervan | | x | x | | | | | |
| Communicatie | | | | | | | | x |
| Curriculum ontwikkeling en -evaluatie | | | | x | | | | x |
| Evaluatie van de leeruitkomsten* | | x | | x | x | x | | |
| Heldere leerdoelen | | | | x | | | | |
| Houding en gedrag van de docent | | | | | x | | x | |
| Leerklimaat* | x | | | | x | | x | |
| Leiderschap* | x | | | x | | | | x |
| Management | x | | | x | | | | |
| Medische kennis en kunde | | | x | | x | | | x |
| Mentorschap* | | | | x | | x | | x |
| Ontwerpen van onderwijs | | x | | | | | | |
| Ontwikkelen en organiseren van onderwijs | | x | | | | | | |
| Prestaties tijdens het instructiemoment* | | x | x | x | x | x | x | |
| Professionele houding | | x | | | | | | |
| Programma ontwikkeling en –implementatie* | x | | | | | x | | x |
| Rolmodel | x | | | | | | | x |
| Significante leerresultaten | | | | | | | | |
| Studentgerichtheid* | x | | | x | x | | x | x |
| Systeem georiënteerd leren | | | | | | | | x |
| Toetsing* | x | x | | x | | x | | |
| Uitvoeren van onderwijs | | x | | | | | | |
| Uren, publicaties en peer reviews | | | | x | | | | |
| Voorbereiding voor instructie* | | | | x | | x | | |
| Zelfontwikkeling* | x | | x | x | | | | x |

*Noot.* De tabel toont de gevonden beoordelingskaders en hun beoordelingscategorieën voor de beoordeling van de kwaliteit van onderwijs. Waar nodig zijn de categorieën vertaald naar het Nederlands. Een x geeft aan dat binnen het genoemde beoordelingskader (bovenste rij van de tabel) deze categorie beschreven wordt.
* Deze categorie maakt onderdeel uit van het ontworpen categorieënoverzicht.

Tabel 5
*De 10 geselecteerde categorieën voor het categorieënoverzicht*

| Categorie |
|---|
| Programmaontwikkeling en -implementatie |
| Voorbereiding voor instructie |
| Leerklimaat |
| Prestaties tijdens instructie |
| Studentgerichtheid |
| Toetsing |
| Evaluatie |
| Zelfontwikkeling |
| Mentorschap |
| Leiderschap |

*Noot.* De tabel toont de 10 categorieën uit het ontwikkelde categorieënoverzicht. Om te komen tot een zo compleet mogelijk categorieënoverzicht zijn elkaar aanvullende categorieën op basis van prevalentie en inhoud in consensus binnen het vaste onderzoeksteam gekozen. De categorieën van dit categorieënoverzicht zijn uitsluitend bedoeld als hulpmiddel in het rubriceren van de in de literatuur gevonden criteria en vragen op het gebied van de beoordeling van de kwaliteit van onderwijs.

Van de 23 benaderde vakgroepen waren 17 vakgroepen, 11 klinische en zes preklinische vakgroepen, bereid deel te nemen aan het validatieproces. Alle benaderde docenten ($N$ = 37, leeftijd 29-62 jaar, 40% mannen, 60% vrouwen) waren bereid hun onderwijs te laten beoordelen. De beoordeling van het onderwijs werd uitgevoerd door 195 unieke collega-docenten (leeftijd 25-66 jaar, 40% mannen, 60% vrouwen). Zij verzorgden in totaal 387 ingevulde vragenlijsten, verdeeld over 37 unieke hoorcolleges. Het aantal ingevulde vragenlijsten voldeed daarmee aan het door Field (2013) gestelde minimale aantal van 300 vragenlijsten.

Onderzoek naar de verdeling, correlaties, *multicollinearity*, aannames voor een normaalverdeling en steekproefgrootte van de verkregen data, leverde geen bijzonderheden op (Field, 2013). De verdere analyse met behulp van PAF, Direct Oblimin, eigenwaarden > 1,0, 25 iteraties en paarsgewijze exclusie, leverde drie factoren op, maar toonde een screeplot met een buigpunt bij twee en vier factoren. Om deze reden werd de analyse uitgebreid naar twee en vier mogelijke factoren. Analyse naar vier factoren (PAF, Direct Oblimin, vier factoren, 25 iteraties, paarsgewijze exclusie) toonde een aanvullende factor aan met een aanvullende verklarende variantie onder 3%. Deze extra factor voldeed daarmee niet aan de gestelde criteria. Analyse naar twee factoren (PAF, Direct Oblimin, twee factoren, paarsgewijze exclusie) gaf een zevental vragen die niet of dubbel scoorden op beide factoren, met een totale verklaarde variantie van de factoren van 38%. De eerder berekende drie factoren verklaarden 45.8% van de variantie in de vragenlijst. Omdat wij de verklaarde variantie van de vragenlijst bij twee factoren te laag vonden en het aantal vragen met geen of dubbele factorladingen te hoog, werd deze variant afgewezen.

De negatief gestelde vragen (vraag 6, 9 en 18) toonden een zeer lage of negatieve inter-itemcorrelaties met de andere vragen. Omdat vermoed werd dat deelnemers deze vragen niet goed hadden gelezen werden deze vragen verwijderd uit de analyse. Na verwijdering van deze vragen resteerden 18 vragen in de lijst en werd de PAF (Direct Oblimin, drie factoren, 25 iteraties, paarsgewijze exclusie) herhaald. De drie factoren verklaarden 51,0% van de variantie in de vragenlijst. De vragen 11 en 20 scoorden op meerdere factoren ladingen boven ,3 en vraag 16 scoorden op geen van de factoren een lading boven ,3. Verwijdering van vragen uit de gevonden factoren zou geen verbetering van de Cronbach's α opleveren. Na verwijdering van de vragen 11, 16 en 20 werd dezelfde PAF-analyse herhaald. Binnen deze vragenlijst van 15 items konden drie factoren worden onderscheiden met Cronbach's α's van respectievelijk (a) .74 (vragen 1, 2, 3, 4, en 21, eigenwaarde 1.69, verklaarde variantie 11.3%), (b) .82 (vragen 5, 10, 13, 14, 15, en 17, eigenwaarde 5.10, verklaarde variantie 34.0%) en (c) .67 (vragen 7, 8, 12, en 19, eigenwaarde 1.32, verklaarde variantie 8.8%).

In het terugbrengen van de lengte van de vragenlijst van 15 naar de gewenste 10 items werd in de discussie vraag 1 verwijderd omdat deze als min of meer gelijk aan vraag 2 werd beschouwd. Vraag 3 werd verwijderd omdat de details van toetsing al beschreven worden in de curriculumbeschrijving binnen ons UMC, welke voor iedere student en docent online beschikbaar is. Daarnaast werd vraag 3 door slechts 34,7% van de beoordelaars ingevuld. Vraag 5 werd als niet relevant beschouwd omdat al het onderwijs binnen ons UMC door correct Nederlands sprekende docenten in het Nederlands plaatsvindt. Gezien de doelstelling van de opleiding, het opleiden om zorg te kunnen verlenen aan in hoofdzaak Nederlandstalige patiënten, zal dat voorlopig niet veranderen. De vraag werd daarnaast ook altijd als (zeer) positief beantwoord en toonde daarmee in de beantwoording te weinig aanvullende onderscheidende waarde binnen onze praktijksituatie. Vraag 13 werd als min of meer gelijk aan vraag 14 beschouwd en kon naar mening van deze discussiegroep ook worden verwijderd. Vraag 21 werd als minder relevant gevonden omdat de uitkomst van deze vraag als minder bepalend voor de kwaliteit van het onderwijs werd beschouwd. Uit de uiteindelijke lijst van 10 items (zie Figuur 3) konden wederom drie factoren geëxtraheerd worden. Door de deelnemers van de discussie binnen de laatst gevormde discussiegroep kregen deze factoren de constructen: (a) context van het hoorcollege (Cronbach's α ,61, vragen 1 en 2[5], eigenwaarde 1,00, verklaarde variantie 10,3%); (b) leeromgeving (Cronbach's α ,67, vragen 3, 4, 6 en 10, eigenwaarde 1,16, verklaarde variantie 11,6%); en (c) studentgerichtheid (Cronbach's α ,78, vragen 5, 7, 8 en 9, eigenwaarde 3,90, verklaarde variantie 39,0%). De drie factoren verklaarden 60,9% van de variantie in de definitieve vragenlijst.

[5] Na hernummering van de vragen in de definitieve vragenlijst.

| Datum: | Docent: | mee oneens / eens | | | | |
|---|---|---|---|---|---|---|
| | **Stelling:** | -- | - | 0 | + | ++ |
| 1 | Bij aanvang is de docent helder over de leerdoelen van het onderwijs | ☐ | ☐ | ☐ | ☐ | ☐ |
| 2 | Om een overgang te maken naar de nieuw te leren stof activeert de docent waar nodig eerst de benodigde voorkennis | ☐ | ☐ | ☐ | ☐ | ☐ |
| 3 | Er worden door de docent verschillende visies op de problematiek aangereikt | ☐ | ☐ | ☐ | ☐ | ☐ |
| 4 | Met goede verhelderende voorbeelden uit de praktijk illustreert de docent de essenties van de leerstof | ☐ | ☐ | ☐ | ☐ | ☐ |
| 5 | De docent onderzoekt regelmatig of studenten het onderwijs begrijpen | ☐ | ☐ | ☐ | ☐ | ☐ |
| 6 | De docent creëert een veilige leeromgeving | ☐ | ☐ | ☐ | ☐ | ☐ |
| 7 | De docent houdt studenten actief betrokken bij het onderwijs | ☐ | ☐ | ☐ | ☐ | ☐ |
| 8 | De docent stimuleert samenwerkend leren en onderling overleg over de lesstof | ☐ | ☐ | ☐ | ☐ | ☐ |
| 9 | Studenten worden door de docent tot zelfstandig nadenken gestimuleerd | ☐ | ☐ | ☐ | ☐ | ☐ |
| 10 | Naar alle individuele studenten, patiënten en collega's is de docent respectvol | ☐ | ☐ | ☐ | ☐ | ☐ |
| Aanvullende opmerkingen en/of voorbeelden bij de beantwoording van de vragen: | | | | | | |

*Figuur 3.* De definitieve vragenlijst.
De figuur toont de definitieve vragenlijst, na analyse en discussie. Door hernummering van de vragen komen de vraagnummers in de definitieve vragenlijst niet meer overeen met de vraagnummers zoals die in de initiële vragenlijst waren vermeld.

## Discussie en Conclusie

### Samenvatting en Beantwoording van de Onderzoeksvragen

Een vragenlijst werd ontwikkeld ter ondersteuning van een vrijwillige, formatieve peer review van kwaliteit van hoorcolleges in academisch medisch onderwijs binnen ons UMC. De centrale vraagstelling voor het onderzoek was: Hoe kan het geven van een peer review van de kwaliteit van een hoorcollege in onze academisch-medische opleiding worden ondersteund met behulp van een korte vragenlijst van maximaal 10 items? Deelvragen bij dit onderzoek waren: (a) Wat zijn de verschillende factoren die uit deze vragenlijst kunnen worden afgeleid? (b) Wat is de interne betrouwbaarheid van deze factoren binnen deze vragenlijst? en (c) Hoe verhouden de uitkomsten van deze vragen zich tot de bestaande literatuur?

De definitieve vragenlijst omvat, conform de opdracht, 10 vragen (zie Figuur 3). De drie gevonden factoren hebben Cronbach's α's van (a) ,61 (context van het hoorcollege, vragen 1 en 2, eigenwaarde 1,00, verklaarde variantie 10,3%), (b) ,67 (leeromgeving, vragen 3, 4, 6 en 10, eigenwaarde 1,16, verklaarde variantie 11,6%), en (c) ,78 (studentgerichtheid, vragen 5, 7, 8 en 9,

eigenwaarde 3,90, verklaarde variantie 39,0%). De factoren verklaren 60,9% van de variantie van de definitieve vragenlijst.

Wanneer in een vragenlijst meerdere factoren kunnen worden onderscheiden is de waarde van de Cronbach's α van de gehele vragenlijst niet relevant (Field, 2013, p. 709). In dergelijke situaties zijn de waarden van de Cronbach's α's van de afzonderlijke factoren belangrijker als graadmeter voor de interne betrouwbaarheid. De waarden van de Cronbach's α's voor de eerste twee factoren, context van het hoorcollege en leeromgeving, waren onder de veelal aanbevolen waarde van ,7 (Field, 2013, p. 709). Feit is echter dat de Cronbach's α beïnvloed wordt door het aantal vragen dat deel uitmaakt van een factor (Field, 2013, p. 709), en binnen deze factoren is er sprake van de aanwezigheid van slechts twee, respectievelijk vier vragen. Met de genoemde scores kan er dus wel degelijk sprake zijn van een goede interne betrouwbaarheid omdat het aantal vragen binnen deze factoren zeer beperkt is. De Cronbach's α waarde van de factor studentgerichtheid, met ook slechts vier vragen, geeft een goede interne betrouwbaarheid van deze factor aan.

Uitgaande van het feit dat de vragenlijst van Newman et al. (2009) de enige gevonden vragenlijst was op het terrein van het ondersteunen van peer review van academisch-medische hoorcolleges werd onderzoek verricht naar de overeenkomsten tussen de definitieve vragenlijst en die van Newman et al. De 10 vragen van de definitieve vragenlijst corresponderen min of meer met drie vragen uit de vragenlijst van Newman et al. uit 2009, maar zijn veelal niet een op een met elkaar te vergelijken (zie Tabel 6). De zeven aanvullende vragen in onze vragenlijst behandelen onderwerpen als voorkennis, het leerklimaat en reflectie, onderwerpen die niet worden beoordeeld met de vragenlijst van Newman et al. (2009). Twee vragen uit de vragenlijst van Newman et al. (2009) die wel in onze initiële vragenlijst aanwezig waren, bleken in het validatieproces onvoldoende onderscheidend te zijn om deze in de definitieve vragenlijst te kunnen behouden.

De inhoud van de uiteindelijke vragenlijst is ook vergeleken met de vragen en criteria in andere gebruikte publicaties. In de criterialijst van Simendinger et al. (2017) staan een aantal criteria verwoord die de vragen 1, 4, 6, 7, 9 en 10 min of meer verwoorden, maar niet elk item van deze criterialijst komt een op een overeen met de items in onze vragenlijst. Dit is verklaarbaar omdat de criterialijst van Simendinger et al. (2007) een overzicht geeft van verschillen in waardering van criteria voor onderwijs in bedrijfskundige scholen in Colombia, Frankrijk, Libanon, Zweden en de Verenigde Staten. Daarmee geeft de lijst een overzicht van criteria binnen andere landen dan Nederland, voor een andere opleiding dan de academisch-medische opleiding. Verschillen tussen de lijsten zouden daarmee mogelijk verklaard kunnen worden door verschillen in cultuur, en opzet van de hoorcolleges, waarbij de hoorcolleges van ons UMC zich onderscheiden van de bedrijfskundige hoorcolleges omdat daar vaak een interactie tussen een of meerdere patiënten en de studenten plaatsvindt.

Tabel 6
*De definitieve vragenlijst in relatie tot de vragenlijst van Newman et al. (2009)*

| De ontwikkelde vragenlijst | De vragenlijst van Newman et al. |
|---|---|
| Bij aanvang is de docent helder over de leerdoelen van het onderwijs | Clearly states goals to the talk |
| Om een overgang te maken naar de nieuw te leren stof activeert de docent waar nodig eerst de benodigde voorkennis | |
| Er worden door de docent verschillende visies op de problematiek aangereikt | |
| Met goede verhelderende voorbeelden uit de praktijk illustreert de docent de essenties van de leerstof | |
| De docent onderzoekt regelmatig of de studenten het onderwijs begrijpen | Monitors audience's understanding of material and responds accordingly |
| De docent stimuleert samenwerkend leren en onderling overleg over de lesstof | |
| De docent creëert een veilige leeromgeving | |
| De docent houdt studenten actief betrokken bij het onderwijs | Encourages appropriate audience interaction |
| Studenten worden door de docent tot zelfstandig nadenken gestimuleerd | |
| Naar alle individuele studenten, patiënten en collega's is de docent respectvol | |
| | Explains and summarizes key concepts |
| | Demonstrates command of the subject matter |
| | Presents material in a clear, organized fashion |
| | Communicates or demonstrates importance of the lecture's topic(s) |
| | Shows enthusiasm for the topic* |
| | Audio and/or visual aids reinforce the content effectively |
| | Voice is clear* and audio-visuals are audible/legible |
| | Provides a conclusion to the talk |

*Noot.* De tabel toont de vragen uit de definitieve vragenlijst en de vragen uit de vragenlijst van Newman et al. (2009). Waar mogelijk zijn min of meer op elkaar lijkende vragen naast elkaar gezet, maar veelal zijn de vragen uit de verschillende vragenlijsten niet een op een met elkaar te vergelijken.
* Vragen uit de lijst van Newman et al. (2009) die in het validatieproces van onze vragenlijst kwamen te vervallen.

In de criterialijst van Valiee et al. (2016) worden eigenlijk alleen de vragen 1, 9 en 10 teruggezien. Deze criterialijst is opgesteld voor de opleiding tot verpleegkundige of verloskundige in Koerdistan. De verschillen tussen de lijsten kunnen ook in deze vergelijking mogelijk worden verklaard door verschillen in de cultuur en de opleiding.

Tot slot vertoont de criterialijst van Pettit et al. (2014) van de op zichzelf staande criterialijsten, de lijsten zonder bijbehorend beoordelingskader, de minste overeenkomst met onze vragenlijst. Alleen vraag 10 uit onze lijst is terug te vinden in de criteria van Pettit et al. (2014). Mogelijke oorzaak hiervoor is dat de criterialijst van Pettit et al. (2014) is samengesteld op basis van de meningen van vierdejaars studenten geneeskunde in de Verenigde Staten. De lijst van Pettit et al. (2014) is daarmee gebaseerd op de meningen binnen een andere populatie dan die van professionals waar onze lijst in hoofdzaak op gebaseerd is.

De inhoud van de definitieve vragenlijst is ook vergeleken met de criteria zoals die worden beschreven in de onderzochte beoordelingskaders. In de verschillende documenten over de BKO (Bestuursstaf, 2012; Van de Wiel et al., 2016) komen, voor de beoordeling van het moment van instructie, de vragen 5, 7, 8, 9 en 10 van de definitieve vragenlijst terug. In de criteria voor de kwaliteit van de voorbereiding van het onderwijs komt aanvullend ook vraag 1 aan de orde. In het Nederlandse beoordelaarskader van Molenaar et al. (2009) komen slechts de vragen 1, 5 en 9 helder terug. De (beperkte) overeenkomst ligt in beide gevallen waarschijnlijk aan het feit dat de genoemde beoordelingskaders meer globaal van opzet zijn en de invulling en weging van details over willen laten aan de lokale praktijk.

Andere overeenkomsten tussen onze vragenlijst en beoordelingskaders zijn te vinden in de internationale literatuur. Binnen het Britse beoordelingskader van de Academy of Educators (2009) zijn de vragen 1, 4, 6, 8, 9 en 10 terug te vinden en binnen het beoordelingskader van de AAMC (Baldwin et al., 2011; Gusic et al., 2014) zijn de vragen 1, 4, 5, 7 en 9 vermeld. Binnen de Stanford Faculty Development Program Tool (Mintz et al., 2015; Owolabi, 2014) worden de items 1, 4, 7, 9 en 10 van onze vragenlijst teruggevonden. Ook bij deze beoordelingskaders geldt dat een specifieke invulling en waardering van de items overgelaten wordt aan lokale interpretatie en weging. Daarnaast zijn de eerstgenoemde beoordelingskaders meer gericht op de waarden van andere, veelal externe beoordelingen, dan op inhoudelijke criteria voor de beoordeling van de kwaliteit van een hoorcollege. Binnen de CanMeds (Frank et al., 2015) worden alleen de vragen 6, 9 en 10 teruggevonden. Dit kan worden verklaard omdat binnen dit kader het verzorgen van onderwijs slechts beperkt aan bod komt. Samengevat komen de vragen uit onze vragenlijst terug in de bestaande literatuur, maar omvat de vragenlijst als geheel een combinatie van vragen die niet eerder in deze samenhang in de literatuur kon worden gevonden. Samengenomen met het feit dat in de literatuur geen Nederlandstalige vragenlijst ter ondersteuning van het geven van peer review van kwaliteit van academisch-medische hoorcolleges kon worden gevonden, maakt de ontworpen vragenlijst een aanvulling op de bestaande literatuur.

De inhoud van de ontwikkelde vragenlijst sluit redelijk aan bij de door Surma, Vanhoyweghen, Sluijsmans, Camp, Muijs, & Kirschner (2019) beschreven 12 bouwstenen voor effectieve didactiek (zie Tabel 7). Vijf van de 10 geformuleerde vragen zijn in deze bouwstenen terug te vinden. Zo reflecteert vraag 2 van de definitieve lijst de eerste bouwsteen, het activeren van voorkennis. Vraag 4 reflecteert het gebruik van voorbeelden (bouwsteen 3), en de vragen 7 en 8 reflecteren in meer of mindere mate het actief verwerken van de leerstof (bouwsteen 5). Vraag 5 sluit aan bij bouwsteen 6, waarin onderzocht wordt of de student het onderwijs begrepen heeft. In de samenstelling van de definitieve vragenlijst komen daarmee vier van de 12 bouwstenen voor effectieve didactiek van Surma et al. (2019) terug.

In de laatste discussieronde, waarin de vragenlijst teruggebracht werd naar een totaal van 10 vragen, werden ook drie constructen geformuleerd bij de drie berekende factoren. De discussie over constructen vond plaats binnen een groep die niet op de hoogte was van het eerder samengestelde categorieënoverzicht dat wij hanteerden voor de indeling van de gevonden vragen uit de literatuur. Waar de vragen uit de initiële vragenlijst oorspronkelijk waren genomen uit de categorieën leerklimaat, prestaties tijdens instructie en studentgerichtheid uit ons categorieënoverzicht, werden nu, zonder kennis van de categorieën bij de deelnemers aan de discussie, de constructen leerklimaat, context van het hoorcollege en studentgerichtheid geformuleerd. Deze verandering werd onder meer veroorzaakt doordat in het validatieproces meerdere vragen uit de categorie prestaties tijdens instructie waren komen te vervallen (vragen 5, 6, 13, 16, 18 en 21). De overgebleven vragen (vraag 2 en 4 uit de definitieve vragenlijst) hadden volgens de discussiegroep een ander gemeenschappelijk construct, namelijk dat van context van het hoorcollege. Deze stelling werd door het vaste onderzoeksteam onderschreven en geaccordeerd.

Het onderzoek had tot doel een vragenlijst te maken voor de ondersteuning van peer review van de kwaliteit van hoorcolleges. Het beoordelen van het moment van instructie is echter slechts een van de vele categorieën waarbinnen onderwijs beoordeeld kan worden. De ontwikkeling en implementatie van een onderwijsprogramma, de voorbereiding van het hoorcollege, leiderschap, mentorschap en zelf-evaluatie zijn voorbeelden van andere complementerende categorieën in het beoordelen van de kwaliteit van onderwijs (Baldwin et al., 2011; Academy of Medical Educators, 2014; Molenaar et al., 2009). Om een breder beeld te krijgen van de kwaliteit van onderwijs van een docent is ook een beoordeling noodzakelijk binnen deze categorieën. Ook dat kan geschieden door middel van peer review, maar daarvoor zijn wel aanvullende vragenlijsten nodig (Van Note Chism, 2007).

Het beoordelen van een hoorcollege door peer review is een van de vele manieren waarop een oordeel gevormd kan worden over de kwaliteit van het hoorcollege. Andere manieren om informatie over de kwaliteit van het hoorcollege in te winnen dienen dan ook gebruikt te worden naast het inzetten van peer review (Berk, 2013; Steinert et al, 2016). Het is van belang dat bij het inzetten van vrijwillige, formatieve peer review gebruik wordt gemaakt van meerdere bronnen, en dat de peer review regelmatig wordt herhaald (Van Note Chism, 2007, pp. 7-8).

De vragenlijst is niet bedoeld als toetsingsmethode, maar als hulpmiddel in de uitvoering van de peer review. Een discussie over de resultaten van de peer review aan de hand van de vragenlijst kan juist bijdragen aan de consensusvorming over wat de beoordeling van de kwaliteit van onderwijs inhoudt. Het gebruik van de vragenlijst kan daarmee een verdere ontwikkeling van onderwijs, in het bijzonder de docentprofessionalisering, stimuleren. Hiervoor is echter wel vervolgonderzoek noodzakelijk om de vragenlijst verder te valideren en te onderzoeken op de inter-rater reliability.

Tabel 7
*De 12 bouwstenen voor effectieve didactiek (Surma et al., 2019) in relatie tot de definitieve vragenlijst*

| Bouwstenen voor effectieve didactiek | | Vragen uit de definitieve vragenlijst | |
|---|---|---|---|
| 1 | Activeer relevante voorkennis | 2 | Om de overgang te maken naar de nieuw te leren stof activeert de docent waar nodig eerst de benodigde voorkennis |
| 2 | Geef duidelijke, gestructureerde en uitdagende instructie | | |
| 3 | Gebruik voorbeelden | 4 | Met goede verhelderende voorbeelden uit de praktijk illustreert de docent de essenties van de leerstof |
| 4 | Combineer woord en beeld | | |
| 5 | Laat leerstof actief verwerken | 7 | De docent houdt studenten actief betrokken bij het onderwijs* |
| | | 8 | De docent stimuleert samenwerkend leren en onderling overleg over de lesstof** |
| 6 | Zoek manieren om te achterhalen of de hele klas het begrepen heeft | 5 | De docent onderzoekt regelmatig of studenten het onderwijs begrijpen |
| 7 | Ondersteun bij moeilijke opdrachten | | |
| 8 | Spreid oefening met leerstof in tijd | | |
| 9 | Zorg voor afwisseling in oefentypes | | |
| 10 | Gebruik toetsing als leer- en oefenstrategie | | |
| 11 | Geef feedback die leerlingen aan het denken zet | | |
| 12 | Leer je leerlingen effectief leren | | |

*Noot.* De tabel toont de 12 bouwstenen voor effectief leren van Surma et al. (2019) in relatie tot vragen uit de definitieve vragenlijst. Niet altijd kan een beschreven vraag uit de vragenlijst een op een met een bouwsteen worden vergeleken, maar sluit de inhoud van de vraag wel aan bij het thema en de strekking van de desbetreffende bouwsteen. Omdat het artikel van Surma et al. na uitvoering van het onderzoek is verschenen is het slechts in de discussie van deze thesis meegenomen.
* Binnen de constructivistische hoorcolleges van ons UMC waar ook patiënten aan deelnemen is een actieve interactie tussen de groep van studenten en de patiënt aanwezig, geleid door de docent.
** In het contact met de patiënt en docent tijdens een constructivistisch hoorcollege met patiënt is er sprake van een samenwerkende leeractie vanuit de studentengroep. Samenwerkend leren leidt tot actieve verwerking van de leerstof en feedback van de docent en medestudenten omdat je onder meer stellingen van anderen tot je moet nemen en eigen stellingen moet verantwoorden.

**Beperkingen van het Onderzoek**

In de samenstelling van de initiële vragenlijst is onder meer gebruik gemaakt van vragen uit wetenschappelijk onderbouwde studentenevaluaties. In de introductie was echter aangegeven dat de uitkomsten van studentenevaluaties onder meer beïnvloed worden door de omstandigheden en persoonlijke, niet-onderwijsgebonden factoren van studenten en docenten (Kamran et al., 2012; Rannelli et al., 2014). In de hoop en veronderstelling dat een aantal van deze verstorende factoren positief kan worden beïnvloed in situaties van een vrijwillige, formatieve, veilige en vertrouwelijke peer review meenden wij deze studentenevaluaties toch te kunnen gebruiken voor de samenstelling van de vragenlijst, omdat de gebruikte studentenevaluaties wetenschappelijk goed onderbouwd waren. De auteur erkent echter dat ook peer review beïnvloed wordt door de omstandigheden en persoonlijke, niet-onderwijsgebonden factoren van de betrokken docent en collega-docenten. Deze problemen worden niet volledig opgelost worden met het gebruik van onze vragenlijst. De auteur is echter wel van mening dat het gebruik van de vragenlijst bij kan dragen aan het structureren van het geven en ontvangen van een peer review. De vragenlijst geeft immers een leidraad aan de peer review, waardoor iedereen naar dezelfde items gaat kijken. In de toekomst moet nog wel onderzocht worden of, en in welke mate, items op gelijke wijze door de beoordelaars geïnterpreteerd worden.

De vragenlijst is ontwikkeld op basis van uitgebreid literatuuronderzoek en een iteratief proces van ontwikkeling en de vragenlijst is uitgebreid getest en bediscussieerd binnen de groep van docenten en onderwijskundigen binnen ons UMC. Desondanks blijft de vragenlijst uitsluitend ontwikkeld binnen ons UMC. Dit sluit weliswaar aan bij de adviezen van de Onderwijsraad (2015) om de beoordeling van de kwaliteit van onderwijs breed en lokaal te ontwikkelen en interpreteren, maar zorgt er ook voor dat de bruikbaarheid van de vragenlijst op andere locaties verminderd kan zijn. Voor het gebruik buiten ons UMC kan het zinvol zijn de vragenlijst breder te valideren, binnen meerdere UMC’s, om de betrouwbaarheid van de vragenlijst te vergroten.

De vragenlijst kan in dit stadium nog niet de wens invullen van de groep van PE’s over het meten van de kwaliteit van hoorcolleges. Hiervoor ontbreken onder meer een tweede validatie en een onderzoek naar de inter-rater reliability. De ontwikkeling van de vragenlijst vormt dan ook slechts de eerste fase van een volledig onderzoek waarmee een vragenlijst kan worden verkregen die peer review van de kwaliteit van hoorcolleges ondersteunt. Verder onderzoek is dan ook nog noodzakelijk om deze vragenlijst in de praktijk en eventuele bestaande kwaliteitskaders te kunnen inbedden.

**Aanbevelingen**

In het validatieproces van de vragenlijst is, op enkele situaties na, gebruik gemaakt van de inzet van collega-docenten uit hetzelfde vakgebied. Deze opzet sloot aan bij de stelling van Siddiqui et al. (2007) dat peer review van hoorcolleges verricht kan worden door collega-docenten uit hetzelfde vakgebied. Het gebruik van collega-docenten uit andere disciplines kan echter een tunnelvisie in de wijze waarop onderwijs wordt verzorgd voorkomen. Daarnaast kan de inzet van collega-docenten uit

andere vakgebieden een afhankelijkheid tussen de docent en de beoordelaar voorkomen (Metcalfe, Farrant, & Farrant, 2010). Wanneer een breder validatieproces van deze vragenlijst wordt uitgevoerd is het wellicht verstandig vaker collega-docenten uit andere vakgebieden het onderwijs te laten beoordelen om deze problemen te voorkomen.

De vragenlijst is getest en gevalideerd vanuit de oorspronkelijke lengte van 21 items. Door wijzigingen in de vragenlijst op basis van de analyses kan niet meer worden gegarandeerd dat de uiteindelijke vragenlijst van 10 items voldoende gevalideerd is. Om deze reden is het noodzakelijk de vragenlijst van 10 items opnieuw te valideren. Dezelfde actie moet worden uitgevoerd voor een Engelstalige versie van deze vragenlijst, om ook deze bruikbaar te maken voor de praktijk.

Omdat met de originele dataset van dit onderzoek, door de aanwezigheid van een shared between level constructie geen inter-rater reliability kon worden berekend is niet duidelijk geworden in hoeverre de beoordelaars van de kwaliteit van de hoorcolleges dezelfde indruk, criteria en mening hadden over de items in de vragenlijst. Een kwalitatief beeld over de interpretatie van de items uit de vragenlijst had mogelijk kunnen worden onderzocht door middel van focusgroepen, waarbij de invullers van de vragenlijst worden bevraagd over hun interpretatie, criteria en waarde van de items. Dit is echter tijdens ons onderzoek niet gebeurd en deze actie is anderhalf jaar na validatie van de vragenlijst niet meer goed mogelijk omdat het vrijwel uitgesloten is dat de samenstelling van de steekproef uit ons onderzoek na anderhalf jaar gelijk is gebleven. Daarnaast kunnen ook de meningen van de deelnemers over de items van de vragenlijst veranderd zijn door een discussie over de kwaliteit van hoorcolleges, waardoor hun huidige interpretatie wellicht niet meer overeenkomt met de interpretatie tijdens het invullen van de vragenlijst. Om deze reden zal in een vervolgonderzoek voor de validatie van de uiteindelijke vragenlijst in de opzet van het onderzoek rekening moeten worden gehouden met het gebruik van focusgroepen of het voorkomen van een shared-between-level-constructie. Om verstorende factoren zo veel mogelijk te beperken, en om een beoordeling zo veel mogelijk te uniformeren, zal daarnaast geïnvesteerd moeten worden in training van en afstemmen in het beoordelen van de kwaliteit van onderwijs en in het gebruik van de vragenlijst (Alabi & Weare, 2014; Berk, 2013; Van Note Chism, 2017, p. 33). Newman et al. (2009) pleitte hierbij voor training van een kleine groep van collega-docenten om uniformiteit in de beoordeling te bevorderen, maar omdat het geven van peer review ook zorgt voor een reflectie op het eigen onderwijs (Thampy, Bourke, & Naran, 2015), adviseert de auteur van deze scriptie juist het trainen van zo veel mogelijk docenten. Met de deelname van een zo groot mogelijk aantal docenten aan peer review wordt de discussie over de kwaliteit van onderwijs breder, en wordt het in een organisatie vanzelfsprekender om een peer review uit te voeren.

Voor het gebruik buiten ons eigen UMC is een bredere validatie van de vragenlijst binnen andere UMC's noodzakelijk. Omdat peer review een goede aanvulling kan zijn op andere beoordelingen van de kwaliteit van hoorcolleges, en een breed gedragen peer review de discussie over de kwaliteit van hoorcolleges kan bevorderen, pleit de auteur dan ook voor de uitvoering van

dergelijke vervolgonderzoeken en trainingen om de ontwikkeling en het gebruik van de vragenlijst, en daarmee de inzet van peer review, compleet te maken.

**Relevantie van de Vragenlijst**

Een Nederlandstalige korte vragenlijst ter ondersteuning van het geven en ontvangen van peer review van kwaliteit van academisch-medische hoorcolleges kon niet worden gevonden in de literatuur. De ontwikkeling van deze vragenlijst kan worden gezien als een eerste aanzet tot de ontwikkeling van een dergelijke vragenlijst. Wanneer de ontwikkeling van de vragenlijst wordt afgerond in een vervolgonderzoek is de vragenlijst een aanvulling op de door Newman et al. (2009) ontwikkelde Amerikaanse vragenlijst. Verdere discussie en onderzoek over de interpretatie van de items van de vragenlijst is echter wel noodzakelijk om dit vervolgproces te ondersteunen.

Een van de doelen van het ontwikkelen van de vragenlijst was een impuls te geven aan de verdere docentprofessionalisering van het academisch-medisch onderwijs binnen ons UMC door peer review te stimuleren. Omdat de vragenlijst na ontwikkeling door meerdere docenten direct in gebruik is genomen kan worden vermoed dat dit doel is behaald en dat een discussie over de kwaliteit van hoorcolleges op gang is gekomen. Een verdere ontwikkeling en het gebruik van de vragenlijst kunnen daarmee bijdragen aan een betere opleiding van artsen, en daarmee betere zorg voor patiënten, in de toekomst.

**Conclusie**

Concluderend is een vragenlijst ontworpen ter ondersteuning van het geven van peer review van kwaliteit van hoorcolleges binnen de academisch-medische opleiding van ons UMC. De vragenlijst bestaat uit 10 vragen en meet drie constructen: context van het hoorcollege, leeromgeving en studentgerichtheid. De huidige vragenlijst dient echter nog te worden vervolmaakt in een vervolgonderzoek waarin opnieuw de vragenlijst, mogelijk op bredere schaal, wordt gevalideerd en onderzoek wordt verricht naar de inter-rater reliability. Wanneer ook dit vervolgonderzoek wordt verricht kan een vragenlijst ontstaan die aanvullend is op de bestaande literatuur.

**Referenties**

Academy of Medical Educators (2014). *Professional Standards* (3rd ed.). Cardiff: Academy of Medical Educators.

Alabi, J., & Weare, W. H. (2014). Peer review of teaching. Best practices for a non-programmatic approach. Communications in Information Literacy, 8(2), 180-191. doi:10.15760/comminfolit.2014.8.2.171

Baldwin, C., Chandran, L., & Gusic, M. (2011). Guidelines for evaluating the educational performance of medical school faculty: Priming a national conversation. *Teaching and Learning in Medicine, 23*(3), 285-297. doi:http://dx.doi.org/10.1080/10401334.2011.586936

Berk, R. A. (2013). Top five flashpoints in the assessment of teaching effectiveness. *Medical Teacher, 35*(1), 15-26. doi:http://dx.doi.org/10.3109/0142159x.2012.732247

Bestuursstaf. (2012). *Universitair kader Basiskwalificatie Onderwijs (BKO) voor wetenschappelijk medewerkers van de Universiteit van Amsterdam.* Amsterdam: Universiteit van Amsterdam.

Creswell, J. W. (2014). *Educational research: Planning, conducting and evaluating quantitative and qualitative research* (4 ed.). Harlow, UK: Pearson Education Limited.

Driscoll, M. P. (2014). *Psychology of Learning for Instruction* (3rd ed.). Harlow, UK: Pearson Education Limited.

Field, A. P. (2013). *Discovering Statistics using IBM SPSS Statistics* (4th ed.). London: Sage.

Finn, G. M., & Garner, J. (2011). Twelve tips for implementing a successful peer assessment. *Medical Teacher, 33*(6), 443-446. doi:10.3109/0142159x.2010.546909

Frank, J. R., Snell, L., & Sherbino, J. (2015). *CanMEDS 2015 Physician Competency Framework.* Ottawa: Royal College of Physicians and Surgeons of Canada.

Gusic, M. E., Baldwin, C. D., Chandran, L., Rose, S., Simpson, D., Strobel, H. W., . . . Fincher, R. M. E. (2014). Evaluating educators using a novel toolbox: Applying rigorous criteria flexibly across institutions. *Academic Medicine, 89*(7), 1006-1011. doi:http://dx.doi.org/10.1097/acm.0000000000000233

Harvey, L., & Green, D. (1993). Defining quality. Assessment & evaluation in higher education, 18(1), 9-26. doi:10.1080/0260293930180102

Hill, M. E., & Herche, J. (2001). Teaching and effectiveness: Another look. Marketing Education Review 11(2), 19-24. doi:10.1080/10528008.2001.11488743

Kamran, A., Zibaei, M., Mirkaimi, K., & Shahnazi, H. (2012). Designing and evaluation of the teaching quality assessment form from the point of view of the Lorestan University of Medical Sciences students-2010. *J Educ Health Promot, 1*, 43. doi:10.4103/2277-9531.104813

Metcalfe, M. J., Farrant, M., & Farrant, J. (2010). Peer review practicalities in clinical medicine. *Adv Med Educ Pract, 1*, 49-52. doi:10.2147/AMEP.S14279

Mintz, M., Southern, D. A., Ghali, W. A., & Ma, I. W. (2015). Validation of the 25-item Stanford Faculty Development Program Tool on Clinical Teaching Effectiveness. *Teaching and Learning in Medicine, 27*(2), 174-181. doi:http://dx.doi.org/10.1080/10401334.2015.1011645

Molenaar, W. M., Zanting, A., Van Beukelen, P., De Grave, W., Baane, J. A., Bustraan, J. A., . . . Vervoorn, J. M. (2009). A framework of teaching competencies across the medical education continuum. *Medical Teacher, 31*(5), 390-396. doi:10.1080/01421590902845881

Newman, L. R., Lown, B. A., Jones, R. N., Johansson, A., & Schwartzstein, R. M. (2009). Developing a peer assessment of lecturing instrument: lessons learned. *Academic Medicine, 84*(8), 1104-1110. doi:10.1097/ACM.0b013e3181ad18f9

NVAO (2018). *Beoordelingskader accreditatiestelstel hoger onderwijs Nederland* (1.0 ed). Den Haag, Nederland: NVAO.

Onderwijsraad (2015). Kwaliteit in het hoger onderwijs. Evenwicht in ruimte, regels en rekenschap. Den Haag, Nederland: Excelsior.

Owolabi, M. O. (2014). Development and psychometric characteristics of a new domain of the stanford faculty development program instrument. *Journal of Continuing Education in the Health Professions, 34*(1), 13-24. doi:10.1002/chp.21213

Pettit, J. E., Axelson, R. D., Ferguson, K. J., & Rosenbaum, M. E. (2014). Assessing effective teaching: what medical students value when developing evaluation instruments. *Academic Medicine, 90*(1), 94-99. doi:10.1097/ACM.0000000000000447

Plomp, T. (2007). Educational design research: an introduction. In T. Plomp & N. Nieveen (Eds.), *An introduction to educational design research* (3e ed) (pp. 9-35). Enschede, Nederland: SLO.

Rannelli, L., Coderre, S., Paget, M., Woloschuk, W., Wright, B., & McLaughlin, K. (2014). How do medical students form impressions of the effectiveness of classroom teachers? *Medical Education, 48*(8), 831-837. doi:http://dx.doi.org/10.1111/medu.12420

Scheerens, J., & Blömeke, S. (2016). Integrating teacher education effectiveness research into educational effectiveness models. *Educational Research Review, 18*, 70-87. doi:10.1016/j.edurev.2016.03.002

Scheerens, J., Luyten, J. W., van Ravens, J., & van Ravens, J. (2010). *Visies op onderwijskwaliteit, met illustratieve gegevens over de kwaliteit van het Nederlandse primair en secundair onderwijs*. Enschede: Universiteit Twente, Vakgroep Onderwijsorganisatie en -management.

Siddiqui, Z. S., Jonas-Dwyer, D., & Carr, S. E. (2007). Twelve tips for peer observation of teaching. *Medical Teacher, 29*(4), 297-300. doi:10.1080/01421590701291451.

Simendinger, E., El-Kassar, A. N., Gonzalez-Perez, M. A., Crawford, J., Thomason, S., Reynet, P., . . . Edwards, J. (2017). Teaching Effectiveness Attributes in Business Schools. *International Journal of Educational Management, 31*(6), 780-800. doi:10.1108/IJEM-05-2016-0108

Srinivasan, M., Li, S. T., Meyers, F. J., Pratt, D. D., Collins, J. B., Braddock, C., . . . Hilty, D. M. (2011). Teaching as a Competency: competencies for medical educators. *Academic Medicine, 86*(10), 1211-1220. doi:10.1097/ACM.0b013e31822c5b9a

Steinert, Y., Mann, K., Anderson, B., Barnett, B. M., Centeno, A., Naismith, L., . . . Dolmans, D. (2016). A systematic review of faculty development initiatives designed to enhance teaching effectiveness: A 10-year update: BEME Guide No. 40. *Medical Teacher, 38*(8), 769-786. doi:10.1080/0142159X.2016.1181851

Surma, T., Vanhoyweghen, K., Sluijsmans, D., Camp, G., Muijs, D., & Kirschner, P. A. (2019). *Wijze lessen, 12 bouwstenen voor effectieve didactiek*. Meppel: Ten Brink uitgevers.

Thampy, H., Bourke, M., & Naran, P. (2015). Peer supported review of teaching: an evaluation. *Education for Primary care, 26*(5), 306-310. doi:10.1080/14739879.2015.1079020

Uttl, B., White, C. A., & Gonzalez, D. W. (2017). Meta-analysis of faculty teaching effectiveness: student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation, 54*, 22-42. doi:http://dx.doi.org/10.1016/j.stueduc.2016.08.007

Valiee, S., Moridi, G., Khaledi, S., & Garibi, F., (2016). Nursing students' perspectives on clinical instructors' effective teaching strategies: a descriptive study. *Nurse Education in Practice, 16*, 258-262. doi:10.1016/j.nepr.2015.09.009

Van de Wiel, M., de Jong, R., Mulder, J., Schlusmans, K. (Eds.). (2016). *De BKO in de praktijk. Inventarisatie en analyse van alle BKO-programma's aan Nederlandse universiteiten.* Deurne: EHON/WUO.

Van Note Chism, N. (2007). Peer review of teaching: A sourcebook (2nd Ed.). San Francisco, Anker Publishing.

*Wet op het hoger onderwijs en wetenschappelijk onderzoek* (1992). Verkregen van http://www.wetten.overheid.nl/BWBR0005682 op 24 september 2019